KTH ROYAL INSTITUTE
OF TECHNOLOGY

Doctoral Thesis in Human-computer interaction

Transparent but incomprehensible

Investigating the relation between transparency, explanations, and usability in automated decision-making

JACOB DEXE

Stockholm, Sweden 2022

Transparent but incomprehensible

Investigating the relation between transparency, explanations, and usability in automated decision-making

JACOB DEXE

Academic Dissertation which, with due permission of the KTH Royal Institute of Technology, is submitted for public defence for the Degree of Doctor of Philosophy on Friday September 16th 2022, at 13:30 in F3, Lindstedtsvägen 26, Stockholm

Doctoral Thesis in Human-computer interaction
KTH Royal Institute of Technology
Stockholm, Sweden 2022

ISBN 978-91-8040-294-1
TRITA-EECS-AVL-2022:44

Printed by: Universitetsservice US-AB, Sweden 2022

Abstract

Transparency is almost always seen as a desirable state of affairs. Governments should be more transparent towards their citizens, and corporations should be more transparent towards both public authorities and their customers. More transparency means more information which citizens can use to make decisions about their daily lives, and with increasing amounts of information in society, those citizens would be able to make more and more choices that align with their preferences. It is just that the story is slightly too good to be true. Instead, citizens are skeptical towards increased data collection, demand harsher transparency requirements and seem to lack both time and ability to properly engage with all the information available.

In this thesis the relation between transparency, explanations and usability is investigated within the context of automated decision-making. Aside from showing the benefits that transparency can have, it shows a wide array of different problems with transparency, and how transparency can be harder to accomplish than most assume. This thesis explores the explanations, which often make up the transparency, and their limitations, developments in automation and algorithmic decisions, as well as how society tends to regulate such things. It then applies these frameworks and investigates how human-computer interaction in general, and usability in particular can help improve how transparency can bring the many benefits it promises.

Four papers are presented that study the topic from various perspectives. Paper I looks at how governments give guidance in achieving competitive advantages with ethical AI, while Paper II studies how insurance professionals view the benefits and limitations of transparency. Paper III and IV both study transparency in practice by use of requests for information according to GDPR. But while Paper III provides a comparative study of GDPR implementation in five countries, Paper IV instead shows and explores how transparency can fail and ponders why.

The thesis concludes by showing that while transparency does indeed have many benefits, it also has limitations. Companies and other actors need to be aware that sometimes transparency is simply not the right solution, and explanations have limitations for both automation and in humans. Transparency as a tool can reach certain goals, but good transparency requires good strategies, active choices and an awareness of what users need.

Keywords: Transparency, explanations, algorithms, automated decision-making, AI, HCI

Sammanfattning

Att något är transparent ses oftast som en önskvärd egenskap. Det offentliga ska vara transparent gentemot medborgaren, och företag ska vara transparenta mot såväl myndigheter som kunder. Mer transparens gör att mer information finns tillgängligt för medborgaren, så att hon kan göra egna och aktiva val i sitt liv, och i takt med att det finns mer och mer information i samhället så kan medborgaren också göra fler val som överensstämmer med hennes preferenser. Tyvärr är det en berättelse som är för bra för att vara sann. Oftare verkar medborgaren vara skeptisk mot ökad insamling av data, hon vill att såväl stat som företag ska bli mer transparenta och hon saknar såväl tid som färdigheter för att verkligen kunna förstå all den information som finns tillgänglig runtom henne.

I denna avhandling undersöks förhållandet mellan transparens, förklaringar och användbarhet, med ett fokus på hur dessa fenomen tar sig uttryck när det rör sig om automatiserade beslut och algoritmer. Utöver att visa vilka fördelar transparens har, visar avhandlingen en mängd problem med transparens, och hur transparens kan vara svårare att omsätta i handling än vad många antar. Den utforska förklaringar, som transparens ofta består av, och dess begränsningar, utvecklingen inom automatisering och algoritmiskt beslutsfattande, samt hur samhället tenderar att reglera sådana fenomen. Avhandlingen använder sedan dessa modeller och tankefigurer för att undersöka hur människa-datorinteraktion i allmänhet och användbarhet i synnerhet kan användas för att förbättra transparens och realisera dess utlovade fördelar.

Fyra studier presenteras som undersöker ämnet från olika perspektiv. Artikel I undersöker hur regeringar och myndigheter använder AI-strategier för att uppnå konkurrensfördelarna med etiskt hållbar AI, medan artikel II studerar hur försäkringsexperter förhåller sig till fördelarna och nackdelarna med transparens. Artiklarna III och IV undersöker båda praktiskt tillämpad transparens genom att begära ut förklaringar av automatiserade beslut, baserat på rättigheter i GDPR. Men, där artikel III jämför implementation i fem olika länder, visar artikel IV i stället hur transparens kan misslyckas och försöker förklara varför.

Avhandlingen avslutas genom att visa att även om transparens visserligen har många fördelar, så finns där också begränsningar. Företag och andra aktörer måste vara medvetna om att transparens kanske inte allt är rätt lösning, och att förklaringar också har begränsad effekt i såväl maskiner som människor. Transparens är ett verktyg som kan användas för att nå vissa mål, men god transparens kräver goda strategier, aktiva val och en medvetenhet om vad användaren vill.

List of Papers

I ***Nordic lights? National AI policies for doing well by doing good***
Jacob Dexe, Ulrik Franke
Journal of Cyberpolicy (2020)

My contribution: The paper was a collaborative effort through and through. The coding, analysis and writing was done in collaboration. However, since I did most of the categorization work (which is seen in the extensive appendix of the paper) I was made first author.

II ***Transparency and insurance professionals: a study of Swedish insurance practice attitudes and future development***
Jacob Dexe, Ulrik Franke, Alexander Rad
Geneva Papers on Risk and Insurance—Issues and Practice (2021)[1]

My contribution: This paper was also a collaborative process. I conducted all interviews together with Alexander Rad who also helped with some of the analysis, and did the rest of the writing and analysis together with Ulrik Franke.

III ***Explaining automated decision-making—A multinational study of the GDPR right to meaningful information***
Jacob Dexe, Ulrik Franke, Kasia Söderlund, Niels van Berkel, Rikke Hagensby Jensen, Nea Lepinkäinen, Juho Vaiste
Geneva Papers on Risk and Insurance—Issues and Practice (2022)

My contribution: I was the project leader for this paper with seven researchers and data collection in five countries. It required co-ordination of research methodology and data samples, as well as the analysis of results. I also wrote the majority of the discussion and conclusions of the paper.

[1]Reprinted by permission from Springer Nature, Palgrave, Copyright 2021.

IV ***Transparency hurdles—investigating explanations of automated decision-making in practice***
Jacob Dexe, Magnus Eriksson, Kristina Knaving
Submitted

My contribution: As with Paper III, I was the project leader for this paper, as well as the main author. The "Hurdles"-section was developed by me after a discussion with Ulrik Franke, and was further improved upon by the co-authors.

Other contributions by the author not included in the thesis:

V ***Towards increased transparency with value sensitive design***
Jacob Dexe, Ulrik Franke, Anneli Avatare-Nöu, Alexander Rad
International Conference on Human-Computer Interaction (2020)

VI ***An Empirical Investigation of the Right to Explanation Under GDPR in Insurance***
Jacob Dexe, Jonas Ledendal, Ulrik Franke
TrustBus 2020: Trust, Privacy and Security in Digital Business (2020)

Acknowledgement

This thesis is (hopefully) a great example of information that is both very transparent and a failure of transparency. It is transparent in that I have attempted to give as much information as possible about the topic at hand, make every argument I can in defense of my position as well as informing you of the weaknesses of that position and arguments against it. At the same time, this thesis is longer than anyone should reasonably be expected to endure in order to get to the point of an argument. The text contains special terminology that very few people will be able to comprehend in full. It spans several scientific disciplines which means I likely fail to do any of them justice. There are arguments that I probably think are very important, from my point of view, but that leave the reader slightly confused. It is written in academic English further narrowing the understandability of my arguments (although it should extend the reach somewhat compared to a thesis written in Swedish). It sets up a series of arguments, but does not come to a particularly satisfying conclusion with a finished recipe for how to actually implement and succeed with transparency.

There are many people to thank for helping me throughout this process. First of all—my supervisor Henrik Artman and co-supervisor Ulrik Franke. Henrik has been a source of experience, of groundedness, and of plenty of joviality. I am not sure how often supervisory meetings are supposed to turn into anecdote sharing sessions or how much supervisor and student are supposed to jokingly insult each other, but at least in this relationship it has been the norm. Henrik has also provided both myself and Ulrik with a connection to reality that we tend to forget. To have realistic expectations on processes, bureaucracy and on the work itself. And to make it fun.

Ulrik, on the other hand, has been the source of ambition, drive and sheer ability. In a very real sense, I would not have been able to complete this thesis if it was not for Ulrik. He framed the project through which I was able to begin my studies, he has served as project manager throughout the four years and over the final six months of writing we had almost daily conversations and check-ins. Ulrik has been a part of the initiation of every idea turned into a paper, every argument turned into a revelation and every problem that was eventually solved. This is, at least in my mind, a thesis written in collaboration with Ulrik.

I also want to thank my co-authors for the various papers. Anneli Avatare-

Nöu, Alexander Rad, Jonas Ledendal, Kasia Söderlund, Niels van Berkel, Rikke Hagensby Jensen, Nea Lepinkainen, Juho Vaiste, Magnus Eriksson and Kristina Knaving. Additionally, since it is hard to work for free, I want to acknowledge the funding from Länsförsäkringsgruppens Forsknings- & Utvecklingsfond (agreement no. P4/18) that made sure I could start my PhD-project and Konkurrensverket (agreement no. Dnr 456/202) that made sure I could finish it and expand on the ideas developed previously. Additionally, thanks to RISE Research Institutes of Sweden for the opportunity to be an industrial PhD-student and work full time with my research and to KTH for doing the same.

Reviewers and respondents have given comments on both papers and this thesis and have, almost always, helped improve the text and my thoughts significantly. This is especially true of Ester Appelgren and her comments on my 90-percent manuscript and Stefan Holmlid for the 50-percent manuscript.

There are many people who have contributed to the process by listening to problems, suggesting solutions or just being there. Although they are too many to list, friends and colleagues, three stand out: Joakim Wernberg has given invaluable help and advice throughout, and is, in part, responsible for me even enrolling in a PhD-program. Rebecka Cedering Ångström has been a great support along these four years, despite us promising each other that we would collaborate through the PhD-experience and never ended up taking more than a course together. And Annika Andreasson, who has the same supervisory team and has therefore been able to understand things that noone else could. Thank you!

Even though they may not always understand what it is I do, mostly because of bad explaining on my part, my family has always cheered me on and for that I am truly grateful!

Most importantly, I am sure I will never find the words to sufficiently express my gratitude and thankfulness for Johanna and our lovely pinscher Tyra. Johanna, thank you for always being there with a kind word, a biting (yet loving) remark when I'm being an ass, and for listening to me rambling on for hours. Thank you for bearing with me. And Tyra, thanks for a sympathetic paw to the face when I have a rough day, or an annoyed paw to the face when I have ignored playing with you for too long. Also, to Shani, who brightened up our lives during the short time he lived with us.

And finally, a note on the cover of this thesis. The illustration of gray and white squares is a pattern often used in graphic design to illustrate an area of an image that is supposed to be transparent when being used various graphical interfaces. In that way, the image used to illustrate the thesis is an example of something that is, simultaneously: transparent,[1] nontransparent,[2] transparent,[3] nontransparent,[4] and transparent.[5]

[1] In that it is commonly understood to be an illustration or symbol of transparency.
[2] In that it is a pattern of solid grey and solid white squares, which is not transparent.
[3] In that it is a clear instruction for a computer program to show it as transparent.
[4] In that it requires an explanation in order to be understood as something transparent.
[5] In that it perfectly describes this thesis' perspective on what transparency is.

What was the aim of this thesis, of the information I will try to present in over these pages? Hopefully to increase understanding about what transparency is, what it can do and how to improve it. Will it do that? Probably not on its own. Change will require that you, reading this, take hold of the ideas within the text and spread them in another context. It will require that someone, me or someone else, rephrase the information herein and present it in lectures or at conferences, or in newspapers. A PhD thesis is a terrible way to achieve what I hope to achieve. I still hope it will be enlightening.

Contents

Chapter 1

Background

Imagine you are invited to a dinner with some friends, perhaps at a restaurant. This, as with many things in life, requires that you make some choices. You will likely know if it is for a birthday or just an evening out based on the conversation with the friend that invited you. You need to figure out what to wear, so you look up what kind of place you are going to, or ask your friend if it is a casual or formal place. At the restaurant, you have a look at the menu to see what looks good, and to make sure it is not too pricey. Perhaps you ask the waiter what the special is today, or ask them to explain what is in a certain dish. You take note of what your friends are doing—if they all order fish, you might let that influence your choice as well. At the end of the dinner you check the receipt to make sure it is correct. You then head home, satisfied with a nice dinner.

All throughout the evening you make choices based on the information available to you. Much of the information you likely did not even think that hard about. You knew the level of formality of the dinner because you spoke to your friends before the dinner. The menus displayed a price after each item and a description of what the dish was. The conversation was pleasant because you could gauge the reactions of your friends and made sure to stay off topics that would be sensitive, hurtful or just plain boring. The more you knew about the place, the food, and your friends, the better your ability to navigate the evening. More information meant that you could make decisions where you were likely to be satisfied with the outcome.

In the physical world people are often able to make decisions based the information available to us. There are contexts, environments and social cues that give guidance when interacting with the world around us. In a sense, this is what transparency is: Information presented to you in a form where it enables you to make choices based on your preferences in relation to the information. It was unlikely all the information that could have been available: You did not receive a breakdown of what share of the price for each item that went to cover the rent the restaurant pays. You likely did not receive a personality assessment of the

waiter, nor their CV, and you likely did not have to consider what mine the metal used to make the knife and fork was sourced from. It is also unlikely that such information would make the experience any better. You received, for the most part, the appropriate information for you needed to be able to make an informed choice about a dinner.

Now, imagine instead that you are signing up for a new service online. Trying to find out what conditions apply might, in the best-case scenario, come down to trying to understand brief explanations of what data is collected, how it is stored, and what effects it can have. More likely you are only given a cursory explanation and are then asked to accept a lengthy contract describing all the details of the service in incomprehensible legalese. If you are interested in finding out how the algorithms of the service actually function, you would need technical expertise to understand it. There are few intuitive clues on the site that help you understand what is going on. You have heard that there is a considerable amount of data collection going on online, but it is hard to figure out exactly what is being collected or from where. The information available to you is likely framed in positive terms, highlighting all the good that comes from using the service and all the neat features it offers, and no word on the potential risks. In the end, the benefits of the service are apparent and the risks are opaque.

I am not arguing that this is what every digital service is like, or that information is always immediately available or clear in the physical world. However, these examples illustrate what this thesis investigates—namely, why it is so difficult to get decent explanations of how algorithms online use personal data and understand why transparency seems to be so tricky in digital spaces. The aim is to increase awareness of how to think about the choices involved in both using and developing algorithmic systems, and improve transparency in such systems to enable people to make more informed decisions.

Improving the ability of individuals to make informed choices is not just a nice-to-have option for organizations in a digital environment. It is, increasingly, becoming a necessity to keep customers and users as they, in turn, are becoming increasingly skeptical towards how data is used online. Indeed, as The Economist noted in 2017, data is now one of the most valuable commodities in the global economy (The Economist, 2017). The increased use of data collection about individuals and automated decisions affects peoples' daily lives. Social media, media consumption, purchasing recommendations, robot vacuums, ride sharing, and the ever present forms of personal information that are filled out for all online purchases all use data and have an impact on people's lives. Nevertheless, these people have little or no agency over that data and its use, nor do they fully understand the situation surrounding it. Hence, they are becoming increasingly reluctant to use such services or products.

The opinion poll Delade meningar (2019, 2020, 2021, 2022)[1] shows that about

[1]Delade Meningar is an opinion poll created by Insight Intelligence together with four partner organizations, different partners each year, focused on consumer attitudes towards data collection

half of Swedes (2019: 42%, 2020: 49%, 2021: 44%, 2022: 43%) are concerned that personal information shared online can be used for purposes with which people are not comfortable. Even more troubling is that two thirds (67%) of all Swedes are negative towards increased data collection, as opposed to the 13% who are positive (Delade meningar, 2022). Morey et al. (2015) also show that customers are concerned about data use, and Appelgren and Leckner (2016) demonstrate that a majority of Swedish consumers try to limit personal information shared online. For most alternatives sampled in Delade meningar (2021) concerning whether people see more advantages or more risks with sharing information online, more people saw risks than benefits, and for half the alternatives (smart home applications, auto-saving passwords on websites, websites remembering payment information, and use of social media) a majority saw more risks than benefits. As an illustrative example, Apple decided to allow users to stop all apps from tracking their phone. Since it was made easy to stop tracking, many chose to do so. This may in turn have contributed to a large fall in Facebook stock prices—since tracking information across platforms is one of the foundations of Facebook's business model (Larsson, 2022).

There is also growing concern about the impact of AI and automated decision-making on human lives. In *Weapons of Math Destruction* Cathy O'Neil (2016) argues that big data and algorithms reinforce inequalities and harm vulnerable communities. An inherent problem is that the algorithms (or mathematical tools) are opaque. O'Neil asks, in the words of math teacher Sandra Bax: "How do you justify evaluating people by a measure for which you are unable to provide explanation?"(O'Neil, 2016, p. 8). Zuboff (2018) argues that algorithms are at the very foundations of the (ominously labeled) *Age of Surveillance Capitalism*, and Velkova and Kaun (2021) describe ways for people to resist algorithmic power.

As this thesis will show, however, despite the many good things transparency brings it is hardly a panacea for the problems facing digitization. It remains unclear what organizations are supposed to be transparent with, and for what reasons. It is also unclear how to explain what happens in a way that people can understand. organizations use technologies that are, at best, difficult to explain; at worst they are unexplainable. Moreover, the legislation meant to direct the transparency efforts and increase people's ability to make informed choices lacks clarity, best practice and proper guidance for implementation. It is also worth considering that, sometimes, it might even be best *not* to be transparent. All these problems, and more, affect the ability for transparency to realize the benefits many assume it has.

Despite these concerns, transparency is a promising tool. The more people know about the choices they make, the better those choices can be. Furthermore, the more people know about the choices they make, the more they can avoid the choices that do not align with their preferences or give them sufficient benefits,

online. I have participated as an independent expert since the beginning (2015-2022) and provided domain expertise as well as continuity to the team.

but more knowledge is needed to figure out exactly how transparency can realize these benefits, in what ways transparency might fail, and how digital services can use transparency as a tool.

It is important to remember, throughout this thesis, that many people want to use digital services and tools to enhance their lives. Most popular services become popular because they provide benefits that outweigh the apparent costs. People want to improve those processes that take up too much time or effort in their lives, and those systems ought to be fair and accountable. In order to accomplish these things there is often a need to process personal data. As shown above, many are reluctant towards such processing (Delade meningar, 2021), but it is also necessary to be aware that the use of personal data in algorithmic decision-making has benefits. This thesis is hopefully a step towards creating a greater ability for customers and citizens to gauge what kind of processing their data is subject to, and to be able to make informed decisions about it.

1.1 Sunlight and data

Transparency legislation has been a tool for society in general, and legislation in particular, for a long time. Sweden adopted the "Freedom of the Press Act" in 1766, which made it possible to gain access to documents, letters, investigations and decisions from public authorities and governing bodies (TF, 1766). In 1913, future US Supreme Court Justice Louis D. Brandeis talked about transparency in the following way:

> Publicity is justly commended as a remedy for social and industrial diseases. Sunlight is said to be the best disinfectants; electric light the most efficient policeman. (Brandeis, 1913)

Brandeis was the first to use the sunlight analogy, which has become synonymous with transparency in the USA. For instance, a piece of legislation that requires that meetings held by government agencies are open to the public is named 'The Government in the Sunshine Act'. Brandeis' argument was that transparency, the light shone on public records, and publicity in financial affairs would be a way to remedy diseases of society, such as corruption. The publicity of financial affairs was designed to serve two purposes: Shame bankers into offering more reasonable terms, and make the market function more efficiently (Lessig, 2009).

It is not only in the cases above where transparency is seen as important in fighting corruption and combating other financial wrongdoing. Non-governmental organizations like Transparency International work to increase accountability through transparency, and the World Values Survey use assessments of institutional transparency as one of their metrics in their global mapping. Looking specifically at digitization and the use of algorithms and AI, Fjeld et al. (2020) analyze thirty six different AI ethics guidelines, and among the eight themes identified, Transparency and Explainability are present in 94 percent of the documents.

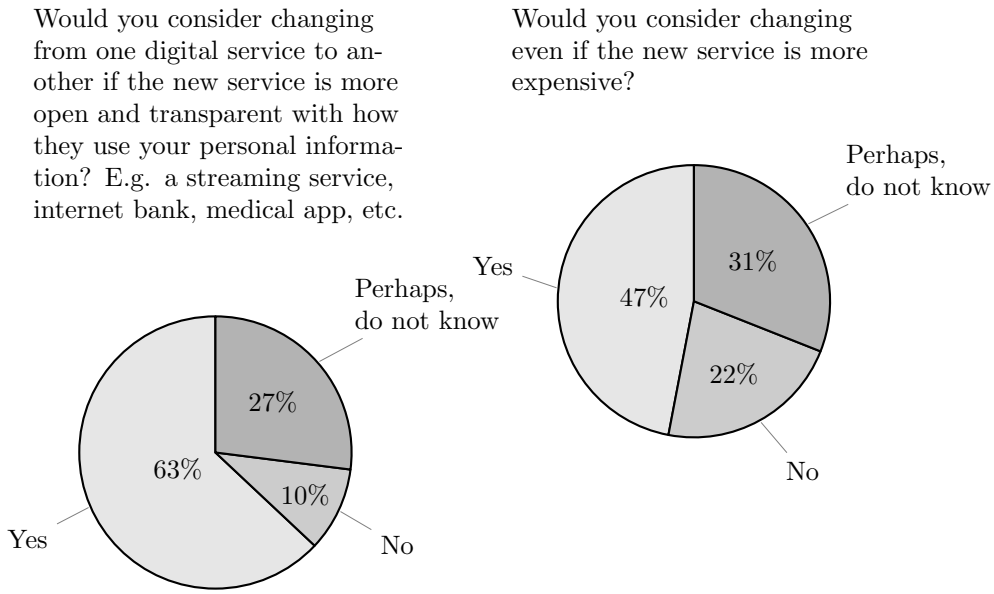

Figure 1.1: Consumer attitudes on openness and transparency (Delade meningar, 2019), translation from Swedish. n = 1000

Transparency, and associated concepts such as accountability, also feature in the Nordic AI strategies analyzed in Paper I. Larsson (2020) also explores this.

Bauhr and Grimes give an overview of the role transparency plays in government accountability, and cite the many ways in which academics have investigated the benefits of transparency. These include: Improving government quality, accountability, dealing with corruption, and stimulating economic growth (Bauhr and Grimes, 2014). For more information on transparency's role in governance, see Hood and Heald (2006).

It is not only governments and public agencies that should be transparent, however. Companies and NGOs also must inform interested parties about what they do, and why. Partly through demands from various legislation, and partly because people are unlikely to engage with completely opaque organizations. As can be seen in Fig. 1.1, when Swedish consumers were asked, in the aforementioned Delade meningar (2019) opinion poll, whether they would be willing to switch from one service to another if the new service was more transparent with how they processed data, 63% answered affirmatively. Out of those 63%, almost half (47%) would consider doing so even if the new service cost more. One reason for this is likely that 2/3 (67%) of Swedes are skeptical towards increased data collection online (Delade meningar, 2021, 2022). Even more damning—Delade Meningar 2022 showed that only 13% of Swedes have a generally positive attitude towards increased data collection online (Delade meningar, 2022).

Indeed, the need and desire for transparency is apparently unmet. Public opinion sees transparency as increasingly important, and legislative measures to increase transparency in technological applications have been introduced (for instance, GDPR (2016)). Technology giants like Facebook and Google have their own transparency reporting regarding how governments are requesting information about users on their services, as do many others. Fleischmann and Wallace (2005) argue that transparency is necessary for humans to be able to maintain autonomy and counter the imbalance of power between developers and users. This involves not just transparency in data, but in assumptions made about reality, elements of the model, and documentation about how the model is built. Hence, Scott (2004) argues that there is a need for transparency and responsiveness in dealing with customer requests about how data is used and decisions are made.

Nevertheless, transparency can also be counterproductive. The idea that 'sunlight is the best disinfectant' might be true, but Lessig (2009) adds to the analogy by saying: "Sunlight may be a great disinfectant. But as anyone who has ever waded through a swamp knows, it has other effects as well." (Lessig, 2009, p. 44). Transparency with the wrong type of information, or even the wrong explanation of the right information, can corrupt the benefits that might be gained from transparency. This idea will be explored throughout this thesis.

Another important aspect to consider is the relationship between transparency and trust. Kim and Lee (2012) find that in e-participation, the more transparent participants feel that local government is, the more they trust the local government. Kang and Hustvedt (2014) show that transparency can have significant effects on trust and the general attitudes towards a company. Kim and Kim (2017) show that transparency can have an effect on the relation between customer satisfaction and trust, and Bhaduri and Ha-Brookshire (2011) show that transparency about responsible or sustainable practices can increase positive associations for a brand among people who are interested in making 'correct' (e.g. sustainable or ethical) choices. Furthermore, Cambier and Poncin (2020) show that signaling transparency (e.g. showing in ads that a company is transparent with its practices) can help the reputation of companies with a poor reputation. However, as mentioned, there are several studies where the impact of transparency is muddled, and there is a lack of consensus on how information transparency should be used by developers and companies.

That is, after all, the aim of this thesis—to understand how transparency works; how the benefits can be realized and the negative effects can be negated.

1.2 Human-Computer Interaction

Before getting into the deeper theoretical foundations of the topic, some time needs to be spent on what scientific domain the thesis is situated in, and how that informs the topic at hand. Human-Computer Interaction is defined by ACM as "a discipline concerned with the design, evaluation and implementation of

interactive computing systems for human use and with the study of major phenomena surrounding them" (Hewett et al., 1992, p. 5).

The field has gone through at least three waves, or paradigms, that all were significant shifts in the domains and interests of study at the forefront of the field (Filimowicz and Tzankova, 2018). The first wave originates in early computing and engineering sciences. The main focus is on how humans and machines interact, and the development and understanding of how to control digital systems. The approach is often pragmatic and heavily focused on practical results (Duarte and Baranauskas, 2016; Filimowicz and Tzankova, 2018; Bødker, 2015).

The development of a second wave was noted by Bannon (1995) and signifies a change in interest towards how humans process information in digital environments, and how digital interfaces create meaning. Bannon chose to title the article "From human factors to human actors", denoting the emerging importance of human abilities and agency. Importantly, the second wave also meant that HCI established itself in a specific domain of study, namely the work environment (Bødker, 2006, 2015; Duarte and Baranauskas, 2016; Bannon, 1995). While the change in attention from interaction with digital interfaces to how digital interfaces function in a work environment seems odd from a 2020s perspective, it does align with how computers were being used at the time. The office computer was becoming more present around 1995, while the home PC was still somewhat of an oddity.

The establishment and recognition of the emergence of the third wave is generally credited to a keynote and article by Bødker (2006). In it Bødker points to an increase in different contexts in which humans interact with digital phenomena, and that those types of applications are no longer confined to the work environment or simple task fulfillment. Instead, the distinction between technologies of work and technologies of home is being blurred, together with a change in the technologies that supply the interactions. The introduction of portable media players and other 'non-productive' devices prompted a reconsideration of what HCI should be and do. It sometimes easy to forget how much technological adaptations have shifted over a relatively short period of time. Bødker states in the article that computers "are increasingly being used in the private and public spheres" (Bødker, 2006, p. 1), a statement that in hindsight seems underwhelming at best. The third wave was, according to (Bødker, 2006, 2015), also a move towards embracing experiences and meaning-making—especially in non-functional and non-rational terms. That is to say, a view of technological interaction that did not emphasize productivity or efficiency (the rational and functional aspects of work) but instead put more attention on aesthetics, emotions and culture.

These waves, Bødker claims, indicate both theoretical and technological shifts. It is abundantly clear that the emergence of new waves correlates with radical shifts in the use of digital interfaces and technologies. There was no focus, in the scientific field of HCI, on leisure or non-functional use of digital technologies before the mid-2000s because such use was still uncommon. There was no clear focus on the use of computers and digital interfaces in the workplace before the early

1990s, because such use was still uncommon. It is also the case that researchers in HCI still work in all the different waves, with modern applications and use-cases. They co-exist, rather than supplant each other, and domains of previous waves continually meet the limitations and technologies of the subsequent waves (Rydenfält and Persson, 2020).

Some scholars are trying to identify the emergence of a fourth wave of HCI. Frauenberger (2019) and Homewood et al. (2020) suggest 'entanglement' of technologies as the main focal point, where the view shifts from a human who uses technologies, to a human who is interconnected and inseparable from her technological artifacts. The mode of study is thus the creation of boundaries between the person and the technologies, and this lifts the technological artifacts as something that has agency and needs to be held accountable. Frauenberger (2019) suggests that through *Entanglement HCI*, HCI should abandon user-centred design, with the user as an object of study, in favor of studying the relationships that technologies enable. Comber et al. (2019) instead suggests 'post-interaction computing', highlighting the fact that while the third wave focused on "designing computing systems for interaction" the fourth wave should focus on "designing interaction for and with computing"(Comber et al., 2019, p.1), that is to say, how humans interact with the underlying systems of the data driven age—algorithms, advertising, large scale data collection and other foundational technologies of the data driven world. Or put slightly differently, HCI should move from looking at the glossy surface to studying the gears underneath. Ashby et al. (2019) claims that the fourth wave should add "politics and values and ethics" on top of the domains, artifacts and methodologies of the previous waves, while Keyes et al. (2019) still claims to belong to the third wave of HCI in a paper proposing an *Anarchist HCI*, which ought to be considered a highly political stance.

Transparency and explanations of automated decision-making algorithms do not fit neatly into any of the waves. Instead, this thesis takes inspiration from each of the waves. From the first wave comes questions on how to produce explanations from automated systems on a technical level, emerging from a technological development that humans are still learning to interact with. From the second wave comes a recognition of the importance of cognitive aspects, where humans are the ones that have to understand the explanations given. It is also a functional perspective—transparency serves a definite purpose, and (often) aims to achieve concrete goals. From the third wave comes a broad adaptation of technologies, ever present in almost every aspect of humans' daily lives. Interaction with technologies cannot be confined to work but comes into the user's life through multiple modes. Finally, from the fourth wave, to the extent that such a thing exists as a cohesive idea, comes the recognition of the entanglement of technologies with human lives, and the need to understand human interactions with the underlying technological infrastructure. In previous waves, technology still appears as an electable, something that humans choose or refuse to interact with. That choice becomes more and more implausible. Humans will likely have to interact with technologies and the decisions of algorithms throughout their

lives. In this way, technology becomes intimate. Not necessarily in that it is close to our body, but rather that it is close to our person—or our digital self. Additionally, from the fourth wave comes the recognition of the importance of values and ethics in the use and development of technologies, and how transparency and explanations play a role in the realization of such aspects.

HCI and Usability

A central concept of HCI, at least in the first and second waves, is usability. According to the International Standard Organisation (ISO), usability is defined as the "extent to which a system, product or service can be used by specified users to achieve specified goals with effectiveness, efficiency and satisfaction in a specified context of use"(ISO 9241-11:2018, definition 3.1) (ISO, 2018). This definition was amended in 2018 to include both 'system' and 'service' (Bevan et al., 2015).

The ISO-definition contains concepts that are worth considering. When thinking about the process of creating software or digital solutions in general, effectiveness and efficiency seem to be central concepts. It is generally preferable for digital solutions to be effective, hopefully more effective than non digital alternatives, and to be efficient, hopefully more efficient than non-digital alternatives. With that in mind, 'satisfaction' is a less obvious goal for a software system, product or service. For a more complete breakdown of how to operationalize usability, see (Speicher, 2015).

Satisfaction, according to ISO, can be defined as the "extent to which the user's physical, cognitive and emotional responses that result from the use of a system, product or service meet the user's needs and expectations"(ISO 9241-11:2018, definition 3.1.14) (ISO, 2018). In fact, one of the reasons why the ISO standard 9241-11 was revised was to clarify that user experience is an important factor for realizing satisfaction when it comes to usability. The previous definition of satisfaction was simply "freedom from discomfort, and positive attitudes towards the use of the product" (Bevan et al., 2015). Previous research has often seen satisfaction as a by-product of achieved effectiveness and efficiency-goals (Hassenzahl, 2001; Bevan, 2010) and not as a goal to be achieved in and of itself. The new definition hopefully changes this.

Norman (1988) argues that designers (in broad terms) should strive to bridge the gulf of evaluation and the gulf of execution. The *gulf of evaluation* is a reflection of "the amount of effort that the person must exert to interpret the physical state of the system and to determine how well the expectations and intentions have been met. The gulf is small when the system provides information about its state in a form that is easy to get, is easy to interpret, and matches the way the person thinks of the system" (Norman, 1988, p. 51). Even if Norman describes physical systems, the same can be said for digital ones. By the *gulf of execution* Norman means the "difference between the intentions and the allowable actions" (ibid.). It can be measured by "how well the system allows the person to

do the intended actions directly, without extra effort" (ibid.). Bridging these two gulfs requires "bringing into a structural alignment designer intentions, the user interface and user mental models" (Bardzell and Bardzell, 2015, p. 43). A user's mental model is the user's understanding of how a product works—the user's expectations. This means that in order to bridge these gulfs it is necessary to try to manage expectations that users have regarding the technological systems.

Any user interacting with a system will come to the situation with a certain set of experiences, prior knowledge and assumptions. Transparency and explanations are a way to make sure that the expectations of the user are matched as closely as possible to the functioning of the system. If they are, the user's expectations are less likely to be violated, and therefore there is a higher chance that the user will be satisfied by the interaction. That is, transparency—used correctly—can be a guiding light for users. As shown in section 1.1 there is evidence of ties between transparency and satisfaction, in part because there are ties between transparency and trust (Kim and Lee, 2012; Kang and Hustvedt, 2014; Kim and Lee, 2012; Bhaduri and Ha-Brookshire, 2011; Cambier and Poncin, 2020).

When bridging the gulfs it is also important for designers to know who is on the other side of the gulf. It is not simply that it is a user, but there needs to be an idea of the particulars of that user—simply put: Who does the designer design for. Often, not knowing who a system is meant to support and which processes it is meant to contribute to, it creates confusion and lack of accountability. (Lindblad-Gidlund et al., 2010, p.44)

Another strand of HCI is, of course, the design of the systems in the first place and the work of designing digital systems in itself. Here, there are efforts to improve transparency in digital systems (Binns et al., 2018; Bove et al., 2022; Alvarado and Waern, 2018; Cheng et al., 2019; Ehsan et al., 2021; Rader et al., 2018) as an important stepping stone to improving user satisfaction with digital products. Bove et al. (2022) is especially noteworthy as they explicitly investigate how transparency affects satisfaction, and find that transparency has a significant effect on that relationship. There are, however, also problems with how much can be explained and how much can be made transparent, which will be presented in subsequent chapters.

Either way, there seems to be an important role for HCI to play in improving transparency—both in that transparency is a necessary factor to achieve satisfaction, and that transparency could be used to increase those feelings.

1.3 Research questions

It seems that in general terms, transparency is almost always seen as something that is intrinsically beneficial. It is almost always something advantageous and desirable. If that is the case—why is transparency so hard to achieve?

Transparency could help make sure that humans can make the best possible decisions about their presence in the digital world. But does transparency work

the way people assume it works? In what ways does transparency fail? And what benefits does transparency bring?

The constituent papers of this thesis all try to investigate different aspects of the questions presented in this Background chapter. As a whole, they try to answer the following research questions:

RQ1 What benefits does information transparency generate?

RQ2 Why is it so hard to achieve transparency in automated decision-making?

RQ3 In what ways does transparency relate to usability?

Chapter 2

Theories, extant research and contexts

This chapter will cover five different areas which all contribute to answering the research questions in the previous chapter. First, the chapter investigates what is actually meant by transparency, and how it works, as well as ways in which transparency fails. Second, it explores explanations, what their role is in transparency and how human decision-making relates to algorithms. Third, there is a brief introduction to what algorithms and artificial intelligence are, laying the groundwork for the things that should be transparent and explained. Fourth is an overview of how regulation works, as well as how the GDPR affects transparency, and fifth, the chapter looks more deeply into what role usability plays in realizing the benefits of transparency and how designers ought to work with transparency going forward.

2.1 Theories of transparency

Turilli and Floridi talk about transparency as "information transparency", partly to differentiate between the transparency that makes something invisible rather than the transparency that makes something visible, and partly to acknowledge what is being made visible. Information transparency, according to the authors, is commonly understood as "the process of making explicitly and openly available (disclosing) some information that can then be exploited by potential users for their decision-making process" (Turilli and Floridi, 2009, p. 105f), especially in academic disciplines such as information management and business ethics. These disciplines fit well into this thesis—it is in the practices of business use of information that transparency, explanations and usability are studied.

Ball (2009), on the other hand, argues that transparency, at least within public policy or administration literature, is best understood through three metaphors: 1) That transparency is a public value that counters corruption; 2) that trans-

parency is akin to open decision-making in government bodies; and 3) that transparency is a complex tool for good governance across different institutions. Ball claims that these metaphors convey how organizations are supposed to conduct their activities—with a mandate given by the public. Kwan et al. (2021) map out how transparency works as a non-functional requirement in design by use of a SIG (Softgoal Interdependency Graph). They present an eco-system of different kinds of transparency, policy positions, ways to package information and ways to access information, all of which correlate to improve transparency. They also show that transparency has a helpful impact on trust in general. Larsson and Heintz (2020) show that in the context of AI, transparency is intimately associated with trust and accountability. They also argue that transparency must be understood in its applied context.

The previous chapter showed examples of where transparency affects trust in different ways (Kim and Lee, 2012; Kang and Hustvedt, 2014; Kim and Kim, 2017; Bhaduri and Ha-Brookshire, 2011; Cambier and Poncin, 2020), but the effects shown have not been described in detail.

Kim and Lee (2012) looked specifically at e-participation (participating in an online environment) in local governments and found that if local residents assess their local government as transparent, their trust in that institution also increases, and the perception of how transparent the local government is is associated with how much citizens feel that they can influence government decision-making. In essence, if they have agency over decisions, they feel like they know more about the system and are more inclined to trust it. Park and Blenkinsopp (2011) find that transparency acts as a moderator between satisfaction and corruption, increasing citizen satisfaction while reducing corruption. Kang and Hustvedt (2014) shows that transparency and social responsibility can have an effect on both trust and attitudes towards a company, including purchase intentions, and that transparency has a larger effect in those relationships than does social responsibility.

Looking instead at how transparency affects purchasing, Bhaduri and Ha-Brookshire (2011) showed that when consumers have a strong desire to make responsible purchases, that increases the value of transparency from a company. However, they also reveal that there is a general distrust towards the legitimacy of claims made. On that note, Cambier and Poncin (2020) show that when brands (or companies) try to show that they are transparent, that may increase the trustworthiness of the brand. The effect is even stronger for brands with a poor reputation. Kim and Kim (2017) show, much like Park and Blenkinsopp (2011), that transparency acts as a moderating force between both corporate social responsibility and corporate ability, which in turn affects both customer satisfaction and trust.

Kizilcec (2016) shows that when expectations are met it does not matter how transparent the system was in terms of trust but for subjects whose expectations were not met, transparency could have a significant impact in restoring trust to levels similar to other subjects. As an example, say that you have submitted an application for funds to an institution. If your application is approved, you are

unlikely to be very disappointed, or feel like you need to question that decision. However, if the application was denied, you would want an explanation to avoid losing faith in the institution. If such an explanation is provided, and is understandable and well motivated, then you are much more likely to still trust the institution, even if your applications was not approved.

Ways that transparency fails

Having shown what transparency is, and how it functions, it is necessary to acknowledge that there are also limitations to transparency. Sometimes it seems difficult to realize the benefits described in the previous section. At other times, it may be that transparency is simply the wrong solution. It could even be that transparency is directly counterproductive.

Lessig (2009) was quoted in section 1.1 saying that even though sunlight can be a good disinfectant, it can have other effects as well. He argues that the move towards more transparency in the public sector may be misguided. The ability of humans to draw conclusions from incomplete data puts such transparency at risk to confirm prejudices rather than enlighten people about realities. An example of this is the often very rudimentary analysis based on information about politicians receiving financial contributions from corporations (this might be especially true for the US context). Such information is often used to argue that the politicians only have the opinions they have because they receive financial contributions, where the contribution is the cause. However, it is likely that the politicians receive a financial contribution from corporations because the politician has views that align with the corporation; in other words, the politician's views are the cause of the contribution. Lessig does not mean that transparency is not important, or that it should not be a necessary requirement on public administration (or companies for that matter). Rather Lessig is saying that there has to be an awareness of negative effects when introducing reforms: "Reformers rarely feel responsible for the bad that their fantastic new reform effects. Their focus is always on the good. The bad is someone else's problem. It may well be asking too much to imagine more than this. But, as shown, the consequences of changes that many see as good, we might wonder whether more good might have been done had more responsibility been in the mix" (Lessig, 2009, p.43-44).

Allan and Berild Lundblad (2021b) argue that when there is incomplete information, where complete information could be considered information about causal reasons for all decisions, then humans tend to fill in the gaps with interpretations that can be worse than actual facts. Humans, in such situations, tend to think about motivations and intentions that are worse than those the actors involved in the decisions actually have.

With respect to algorithmic decisions, de Laat (2018) considers whether full transparency is actually a desirable state of affairs. de Laat focuses especially on the effects on accountability, arguing that the default state is often that algorithmic systems are opaque, and that such a state is generally seen as undesirable for

the public. The desirable state of affairs would be full transparency. However, de Laat argues that there are four adverse effects that make full transparency impossible, leading to the conclusion that full transparency, at least when it comes to achieving accountability, is only feasible with oversight bodies, i.e., organizations that others can put their trust in. What are these adverse effects? First, that it would be detrimental to privacy to have the data sets, on which the machine learning algorithms are built, fully exposed to the public; second, that full transparency concerning the machine learning models would open up those systems to be gamed or taken advantage of by outside actors; third, that most business would consider machine learning algorithms as information that should fall under business confidentiality; and fourth, because the state of technological development means that algorithms tend to be inherently opaque, as will be shown in section 2.3. The conclusion is that if full transparency is not possible towards the public, customers, or competitors nor in the machines themselves, then it should only be necessary towards government oversight agencies. Note that this is specifically for the goal of reaching accountability—other goals might be achieved other ways. The lesson may be that, like Lessig (2009) argues, it is important to think about the possible negative consequences of transparency to realize the best possible solution.

In addition to the many positive relations that transparency has Kwan et al. (2021) also show that *vague transparency, false transparency, assumptions* and *programmed ethics* all affect transparency negatively, and therefore affect trust negatively. As an example, false transparency can be when a company only releases positive information, but fails to disclose negative information. As for assumptions, they point out that if assumptions are made about information that is not transparent, this may have an adverse effect, echoing Allan and Berild Lundblad (2021b) who say that when there are gaps in the information that companies are transparent with, customers tend to fill those gaps with assumptions that are generally worse than what has actually occurred.

In more practical terms, Sørum and Presthus (2020) investigate how companies respond to requests covering the right to access and right to portability (Article 15 & 20) in the GDPR. The authors use their customer status with the companies to be able to send the requests *as consumers*, rather than as researchers. While they did receive some information from a majority of the 15 companies they contacted, almost none shared meaningful information about possible automated decision-making. Sørum and Presthus mostly make note of this, and conclude by stating their concern that the companies may not have properly understood the distinction between the two different GDPR articles, and that customers who are looking for such information should be sure to specifically delineate which articles they are referring to and what kind of information they want access to.

In another example, Appelgren (2017) shows that even if a privacy policy is nominally transparent and open about what data is being processed, the reasons given for that processing can conflate the understanding of, for instance, business

logic and journalistic logic—thereby risking the trust of users.

Pro-ethical condition

In order to figure out what transparency can actually accomplish, it might be worthwhile to look more deeply at the concept and what forces affect it. As mentioned, Turilli and Floridi (2009) describe that by 'information transparency' researchers in business and business ethics often mean "the process of making explicitly and openly available (disclosing) some information that can then be exploited by potential users for their decision-making process" (Turilli and Floridi, 2009, p. 105f). In such a definition, there are several important keywords that should be understood separately, in order to tie together the whole. There is a process of making information available, just as there is a need for the information to exist in a manner that is exploitable (usable). Moreover, there is a need for the information to have an intended target and possibly intended use, and there is a need for the information to say something about something real to make a decision-making process relevant.

It is important to note, as the authors also do, that transparency is not an ethical value in and of itself. Even though it is at the forefront of AI policies (Fjeld et al., 2020; Floridi et al., 2018; Larsson, 2020), Turilli and Floridi note that there are many instances where software or another technical artifact discloses information that does not lean in any one way ethically. For instance, computational processes, account balances and other types of information are disclosed, but hold no ethical value in and of themselves.

Instead, transparency can be considered a *pro-ethical condition*. It is a mechanism that affects ethical values through different means. Many of the ethical values that society tends to want to uphold require that information is disclosed in some form or another. There are two main relationships that transparency has with other ethical values, *Dependence* and *Regulation*, and each, in turn, has two possible effects on the ethical values: *Enabling* or *Impairing*.

In the dependent relationship information transparency is required to endorse or realize the ethical principles. Turilli and Floridi suggest that for the value of accountability to be realized, some information must be disclosed. Similarly, for any consumer to give informed consent in any agreement, they will need access to information about what that agreement entails. Informed consent could not exist without some form of information transparency, as Fleischmann and Wallace (2005) also emphasize. Conversely, certain ethical values instead regulate what type of information transparency should exist— namely, how information can be used in certain context. Privacy can only be maintained if audience that receives access to certain information is limited. Anonymity also works in this way, with even harsher regulation.

Through these relationships information transparency is ethically enabling "when it provides the information necessary for the endorsement of ethical principles (dependence) or [...] when it provides details on how information is con-

strained (regulation)" (Turilli and Floridi, 2009, p. 107). When, on the other hand, the information is false, inadequate or excessive in relation to what is needed, the information has an impairing effect on the ethical values, which is to say—transparency can have a negative impact.

How information is created

One important aspect of the different perspectives on transparency in section 2.1 is the process of making information available. This conjures up two questions: What is information, and how is it created?

Information, Turilli and Floridi (2009) argue, consists of "meaningful, veridical, comprehensible, accessible and useful data". They also posit that these need to contain true (as in truthful) semantic content. This differentiates the information from data, as data is instead said to be, at least on a very basic level, a "lack of uniformity". Data is, for instance, a string of binary digits, or code, or the letters in a book. It can be combined and codified into languages, programs or other types of interactive media.

In order to create information, data needs to be subjected to various types of operations, such as organising, combining or interpreting it. As previously seen, information transparency can have a dependent or regulatory relationship with ethical values. Nevertheless, the creation of information from data also influences various ethical values. Turilli and Floridi show that accuracy, fairness and impartiality are dependent on the process of creating information, yet without knowing something about how the information has been created, those values cannot be realized.

If there is no disclosure concerning how information has been created, it is difficult to accurately judge the ethical implications of the information. Therefore, the information creation process, as in subjecting data to some kind of operation, is a necessary component of information transparency. Turilli and Floridi are adamant that companies, organizations and public institutions "cannot limit their ethical involvement to public declarations of intent". Instead, they must show how the information has been created, what ethical principles they are committed to upholding, and how the ethical principles are prioritized and put into practice.

In short, Turilli and Floridi (2009) mean that information is created from the operations performed on various sets of data. These operations, if transparent, endorse certain ethical principles. The information created can, in turn, be used to either enable or impair ethical principles through a dependent or a regulatory relationship. This process can be seen in Figure 2.1.

Another way to frame this process is to instead look at the aims of the process, what the transparency is meant to achieve. The model presented points out that information is used to endorse or impair ethical values. However, the ethical values are not reached in isolation. As Turilli and Floridi also point out, information transparency (in business and information management research) is often meant to make information visible to an actor who then uses it to 'exploit'

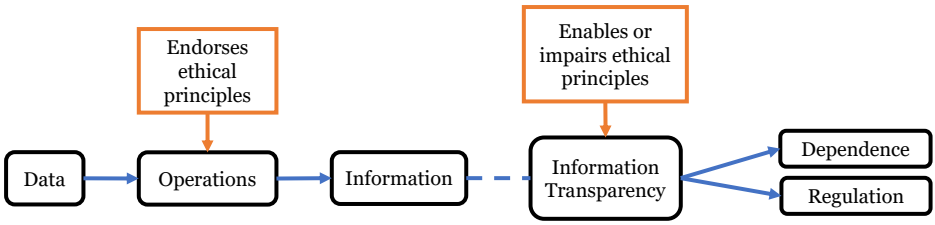

Figure 2.1: Summary of Turilli and Floridi's model of information transparency.

a decision-making process. This actor is only implied in the models Turilli and Floridi present, but for information to be created and presented in such a way that a specific ethical value is achieved, an actor must make a choice. This is shown in Figure 2.2.

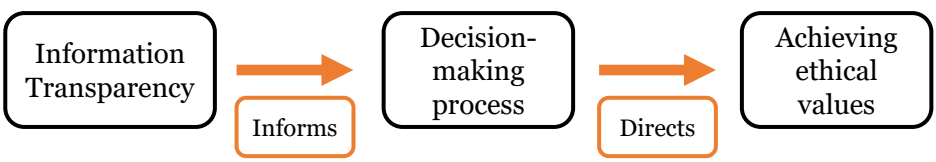

Figure 2.2: A simplified process of decision-making in Turilli and Floridis model of information transparency.

2.2 Explanations

The previous section explained what transparency can do, where it fails and in what ways information transparency is constituted. This section, instead, looks at the applications of transparency relevant for this thesis, namely explaining automated systems. It begins by covering the first part of that statement—explaining. How are digital systems generally explained, or other things for that

matter, and how do explanations tend to work? It will also cover in what ways explanations fail to accomplish the goal of creating understanding. The next section, 2.3, covers the technologies—the automated systems—to gain a sense of where the explanations apply.

The introduction to *Explanations and Understanding* by Keil (2006) reads as follows:

> Humans are driven to acquire and provide explanations. Within months of uttering their first words, children ask "why." Preverbal infants explore phenomena that puzzle them in an attempt to uncover an explanation of why an effect occurred. As adults, we must frequently choose between explanations of why politicians lost, why the economy is failing, or why a war is not winnable. Moreover, explanations are not merely the work of experts. Our friends explain why they have failed to honor a commitment or why a loved one is behaving oddly. Our enemies may offer unflattering explanations of our successes. Explanations are therefore ubiquitous and diverse in nature. (Keil, 2006, p. 228)

Explaining problems, methods and intricate processes is the bread and butter of scientific work. It is what any researcher does any time they produce a paper or give a lecture, or even share a coffee with someone who happens to ask a question. Indeed, any academic theory of science course likely deals with the question of explanations. There are many different types of scientific explanations, but the central model is the Deductive-Nomological model (see, e.g., Woodward and Ross, 2021). This model consists of an *explanandum*, something that is to be explained, and an *explanans*, the statements proposed to explain the thing. As Woodward and Ross state, "the explanation should take the form of a sound deductive argument in which the explanandum follows as a conclusion from the premises in the explanans". In addition, *explanans* must contain some general law of nature or regularity to be valid (Woodward and Ross, 2021). This way of thinking about explanations is foundational for scientific thinking, but perhaps not as intuitive as the consumer-oriented explanations of interest in this thesis.

Keil (2006), as seen in the quote above, as well as Wilson and Keil (1998) take a different approach to explanations than Woodward and Ross (2021). They instead try to look at how everyday explanations work, and how humans make sense of less than perfect explanations. Wilson and Keil (1998) define an explanation as "an apparently successful attempt to increase the understanding of that phenomenon" (p. 139). While Keil does not define the term, he says that explanations are transactional in nature, an attempt to communicate an understanding, often from one person to another, or from one institution to an individual. Humans use explanations for a multitude of reasons. They are used for the prediction of similar events to increase the ability to anticipate them; they are used in diagnosis, either for diseases or for when things go wrong so that the problems can be fixed, or to explain why a certain solution did or did not work. They are even used

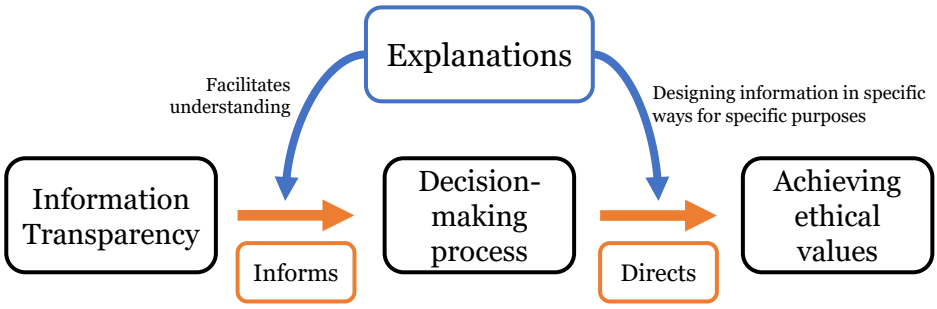

Figure 2.3: A model for transparency and explanations.

for events that are or seem improbable and perhaps only based on randomness. Humans want explanations either to try and learn from the events or to be able to blame a person or an event for some negative consequence. Explanations are also used to justify actions, to make certain decisions appear sensible or appropriate in a certain scenario.

Explanations are, in the real world, often somewhat *shallow*. Explaining every detail of every occurrence would take more time than any person has available to them—a complete explanation of all causal relations that can be explained in any given situation is enormous (Wilson and Keil, 1998). Keil (2006) points out, however, that "explanations don't have to work in real time. Instead, they may frequently serve to help people know how to weigh information or how to allocate attention when approaching a situation" (Keil, 2006, p.234). That is to say, explanations help us make decisions.

Building on the model presented in Figure 2.2, it seems that explanations form an important role in facilitating the decision-making process, which the information transparency is aimed at. The decision-making also must design explanations for consumers so that they, in turn, can understand the information presented (and thereby the ethical values can be realized). Explanations could be considered an instance of 'operation' regarding information (Turilli and Floridi, 2009) but rather than creating information from data, it serves to create understanding from information. This model of transparency is shown in Figure 2.3 and is the model or theory of transparency this thesis builds on.

Section 1.2 showed that satisfaction (as a goal for usability) is in part connected to the mental models of the user (ISO, 2018; Norman, 1988; Bardzell and Bardzell, 2015). Keil also points out a difference between mental models and explanations. Mental models are "readouts of relations from a mental array and are often understood in spatial terms" (p. 229). Explanations are, in contrast, not blueprints or plans that are simply descriptive, they also include the interpre-

tations of those plans. The transactionality of explanations is key here—a mental model is tied to the individual's idea of a thing and is not meant to be transferred to others. In the context of this thesis, the designer's (or product owner's or marketing professional's) effort to transpose their mental model of a system to an explanation that will be shown to consumers is transparency in practice.

Berild Lundblad (2018) argues that when it comes to explanations regarding AI decisions, or algorithmic transparency, it might not be worthwhile to demand a full explanation to verify that the decisions taken are correct. Rather, validation of algorithmic decisions should instead be based on an outcome analysis. That is, by looking at the outcomes of decisions, inferences can be made about biases, the quality of the decision, and the alignment of the values the systems are supposed to uphold. If the decisions are in line with those criteria, then the system can be assumed to be "correct". Such a system, Berild Lundblad argues, would also have the benefit of being more immediately available for judgments from the public and could therefore, in itself, increase transparency even if certain variables are still hidden from public view. London (2019) echoes this point by quoting Aristotle as saying that society must not "demand in all matters alike an explanations of the reason why things are what they are; in some cases it is enough if the fact that they are so is satisfactorily established" (p.18).

A great example of this reasoning can be inferred from Foyer (2015). Working Dogs (Foyer specifically addresses Military Working Dogs, MWDs) perform a number of different tasks in society. Multiple organizations rely on Working Dogs to find people, ordinances and illegal substances. In all these cases, institutions trust the indications made by the dogs not because they understand exactly how the dogs think, or because the dog can say what specific scent it picked up. Instead, institutions rely on the training of the dog, and the fact that the dogs have a good rate of detection. They also tend to try and verify the indications made by the dogs through lab testing or by humans searching the location instead—i.e. they perform an outcome analysis of the decisions made by an opaque automated system.

To be able to determine the relationship between transparency and explanations, in where it is not evident, a more formal description of different explanations is needed. As it happens, Keil (2006) lists four types of explanations:

1. *Causal patterns* are perhaps the most common type, as humans tend to want to find causality in explanations both given and received. Keil distinguishes between four sub-types of causal patterns: common cause, common effect, linear causal chains, and causal homeostasis. In common-cause explanations humans look for a single event that causes different observable effects, e.g., the stress you feel from work causes bad sleep, anxiety and a bad mood. Common-effect explanations instead pay attention to when various causes come together to create an event. Keil points out that historical events are often explained in this way. Linear chains denote explanations where one thing often leads to a chain of events. Finally, causal homeostasis expla-

nations are ways to explain, for instance, the complex balance of natural ecosystems. Many different causes and effects balance each other and create a stable environment.

2. *Explanatory stance* highlights from what perspective an explanation is given—what that stance it is purporting. A mechanical stance considers physical objects on basic level and how they interact. A design stance instead describes the functions and purposes that a system or entity (much like some HCI-research), and an intentional stance focuses the explanations on the beliefs, desires and other things that make agents act in certain ways. Keil gives the example of a diver tucking in their legs. Such a scenario can be explained "in terms of the physics of rotating objects (a mechanical stance), in terms of the purpose of pulling in the limbs close to the body (design stance), and in terms of the beliefs the diver has about her actions and the motivations that drive them (the intentional stance)" (p. 232). Keil (2006) also notes that atypical stances can be used to add insight into systems for pedagogical reasons, through analogies or thought experiments, but such explanations may also distort understanding.

3. *Explanatory domains* are a sort of explanation that hinges on prior knowledge and an intuitive understanding of certain fields. An economist is likely to explain events in terms of the rationality of individual actors, or through market mechanisms. A mechanical engineer, however, might use torques, forces, and friction.

4. *Social- and emotion-laden explanations* stress the motivational factors behind explanations. It does not necessarily affect understanding, but instead lowers or raises thresholds for acceptance. Social context and emotional responses color how people interpret explanations. There are also explanations that are more visceral, for instance some moral stances can be explained through certain principles that the person adheres to, while "such taboos as incest, sacrilege, and torture" (p.233), are more often explained by people's gut reactions.

Another perspective is presented in Wilson and Keil (1998) where the authors try to explain what mechanisms allow humans to understand and accept explanations that are shallow, or at least much shallower than what researchers tend to produce or expect. Rather than list types of explanations, they list mechanisms, or aspects of explanations, that create 'explanatory sense' even if they do not give a precise explanation of how a thing functions:[1] Hence, *explanatory centrality* is a mechanism whereby certain properties are especially important in a certain domain (similar, but not to be confused with explanatory domains above)—for example, the number of limbs to describe different animals, or risk in determining

[1]These aspects have no specific name in Wilson and Keil (1998) but were given names in Paper III for the sake of improved intelligibility and the ability to refer back to them.

insurance premiums. *Causal power* denotes how humans understand that a certain thing tends to behave in certain ways in certain situations—a hammer tends to pound in nails, a stove tends to heat other objects. *Agency and cause* describes the idea that different agents have associated causes for acting (similar to explanatory stances above) that are not transferable to other explanations—feelings and whim for humans, chemical reactions for chemical agents. *Causal patternings* are similar to agency and cause, but this describes situations where several events proceed in patterns, either in chains such as with linear chains above or in other patterns (Wilson and Keil, 1998, p.154). These are not features that make an explanation good but simply features that many shallow explanations share and that can make even shallow explanations understandable.

Breakdowns in explainability

Having looked at how explanations work, and different perspectives on what explanations might be necessary in different environments, this section instead explores what happens when explanations break down.

London (2019) argues against explainability in the specific domain of clinical decision support systems. He argues that in medical science there are an abundance of decisions made on incomplete information. This is not because of malice or laziness, but because such is the state of the knowledge about the human body, diseases and how to remedy them. "As far back as the ancient Greeks, trust has been connected to the ability to explain expert recommendations" London argues, before pointing out that while critics are concerned by the "atheoretical, associationist and opaque" ways in which an advanced AI makes decisions, such is also the nature of many decisions in the medical sciences. In medicine, the theories used by practitioners for both understanding diseases, and how a certain drug works on such diseases, are often "unknown or of uncertain value" and cannot be explained in detail to a patient. Both doctor and patient simply rely on established correlations, rather than mechanical explanations and known causalities.

In a similar vein, Zerilli et al. (2019) argues that there is a double standard regarding how explainability in algorithms is assessed, especially in AI, because requirements for explanations in algorithms are higher than those required from most human decisions. Zerilli et al. goes through multiple examples of different biases humans have, and different instances of those biases being investigated in scientific literature, and examines how those affect transparency. For instance, when humans are asked to justify how decisions are made after the fact, they have a hard time disclosing accurately exactly why a certain decision came to be. The authors then argue that when demands are made for AI to be explainable and understandable, the standard for what level of explainability is desirable is often much higher than what is expected from human decision-making.

As such, it might be reasonable to say that demands on what kinds of explanations humans should expect from artificial intelligences should not be more

detailed than those demanded from fellow humans.

de Laat (2018) vehemently disagrees with this point, however, questioning whether the "dominant trajectory of developing algorithms that become ever more accurate but ever less intelligible" should be accepted. If the choice is between having to design models that help humans explain the decisions an algorithm has made, or if models should be intelligible by design, de Laat (2018) is firmly in the second camp. That is the only way to achieve full accountability, for experts and laymen alike.

Asadabadi et al. (2020) write about *hidden fuzzy information* and its effects on requirements in procurement processes. Specifically, *hidden fuzzy information* makes it difficult to convey the exact meaning of a requirement. It is a poor explanation of what the conveyor of the text wants. Fuzzy information are words and phrases that do not have a specified value—you cannot read it and immediately understand what is required. "I want some food" is fuzzy both in terms of the amount and what kind of food is required. "I want four pancakes" is a much clearer requirement. *Hidden* fuzzy information is information that would further clarify what is actually meant by the requirements, information that is implied rather than explicit. I might only say that I want four pancakes, but actually mean that "I want [to eat] four pancakes [for dinner]". This way to think about what explanations convey is reminiscent of the shallow explanations mentioned in Wilson and Keil (1998).

Continuing in the field of AI, Lakkaraju and Bastani (2020) show that explanations from AI systems can foster unwarranted trust from users, partly because AI systems do not reflect biases accurately, thus explanations can be misleading. While an AI does not possess intentionality in its statements, the problem is in some ways fuzzy information, in that implicit information (the system being biased) is not made explicit (explained) (Asadabadi et al., 2020).

Keil (2006) did also point out that explanations "help people know how to weigh information or how to allocate attention when approaching a situation" (p. 234). This brings to mind a quote attributed to Orson Welles: "I can think of nothing that an audience won't understand. The only problem is to interest them; once they are interested, they understand anything in the world." In addition, Simon (1971) has said that: "In an information-rich world, the wealth of information means a dearth of something else: a scarcity of whatever it is that information consumes. What information consumes is rather obvious: it consumes the attention of its recipients. Hence a wealth of information creates a poverty of attention". As such, it can be assumed that interest and attention are vital components to explanations, both as a cause and an effect.

2.3 (Explainable) Artificial intelligence

As has already been well established, this is not a thesis about artificial intelligence, nor a thesis about algorithms or automated decision-making in general.

However, these are technologies that are of concern to understanding the needs and requirements of and for transparency and explainability in the digital sphere. It is a set of technologies and technological applications that are both increasing in use, and affecting human lives to a greater extent than before. As such, there is a need to set the scene.

First of all, to an extent, both automated decision-making and artificial intelligences are subsets of algorithms. They are large and broad subsets, but subsets none the less. Cormen et al. (2009) define an algorithm as "any well-defined computational procedure that takes some value, or set of values as input and produces some value, or set of values, as output." (p. 5). While others have disagreements (Yanofsky, 2011), it seems an appropriate definition here. Informally, an algorithm could also be described a set of rules for information processing.

People use and make use of algorithms daily. Your email is kept free of spam by algorithms that take the input "this email looks like other emails filtered as spam" and sends them straight to the trash or filters them out completely. Your music app uses algorithms that match the music you have previously listened to (input) and gives you recommendations for new music like it (output). In a very broad sense, your cooking recipes are algorithms which, based on certain ingredients (input) and ways of preparing them (input, or procedure, depending on circumstances), end up giving you a lovely cake (output).

This definition can be added to by saying that algorithms are sensitive to the quality of the input and the accuracy of the computational process. For example, using salt instead of sugar, or broiling instead of baking will give you a very unpleasant cake.

Nevertheless, artificial intelligence, AI, is a surprisingly difficult concept to define, despite its common use in both pop-culture and academic sciences. Simmons and Chappell (1988) define it as "behaviour of a machine which, if a human behaves in the same way, is considered intelligent". Wanting to rid the field of the troubling implications of trying to define intelligence, Dobrev (2003) opts for saying that "AI will be such a program which in an arbitrary world cope not worse than a human". In 2020 the *Journal of Artificial General Intelligence* devoted an entire issue to comments on a working definition of Artificial Intelligence by Wang (2019), the most well established definition in the field (Monett et al., 2020). The definition supplied is as follows:

> The essence of intelligence is the principle of adapting to the environment while working with insufficient knowledge and resources. Accordingly, an intelligent system should rely on finite processing capacity, work in real time, open to unexpected tasks, and learn from experience. This working definition interprets "intelligence" as a form of "relative rationality". (Monett et al., 2020, p. 1)

In popular use AI tends to describe application of particular methods like machine learning (ML), and particularly neural networks. Rebala et al. (2019) argue that

while AI is about creating machines that are intelligent, ML is about creating machines that can learn to perform tasks. Since learning is tied to the understanding of intelligence, ML is also considered one of the few ways in which real AI can be created. Rebala et al. also stress that certain ML technologies such as neural networks create interpretability problems. Finally, "a neural network is a computational system composed of nodes [...] and the connections between these nodes." (Rogers et al., 1992). Neural networks make classifications by allowing the machine to find statistical correlations in a text or set of texts on its own and then ascribe different nodes (neurons) a value (weight) based on those correlations.

Due to this way of computing, it is impossible to know exactly what line of reasoning such an advanced system uses to come to a certain conclusion or decision—which is why they are often called black boxes (Gasser and Almeida, 2017). This, in turn, has given birth to an entire field called Explainable AI, or XAI. For a more in-depth analysis of explainable AI, see Guidotti et al. (2018); Du et al. (2019); Rai (2020); Meske et al. (2020) among others. Within XAI there are several different methods to use to make a less explainable model into one that can be explained more fully, for instance, decision trees (Andrews et al., 1995; Barakat and Bradley, 2010) or heat and salience maps (Samek et al., 2016; Adebayo et al., 2018).

Because these systems use data (either texts, or images, or numbers) to learn about the world, they also run into several problems with the incompleteness of the data they receive. If such a machine is given a bunch of pictures that are labeled as 'dog', it can infer that certain patterns in those pictures are indeed pictures of dogs. It will not, however, understand what a dog is; it can only determine the statistical patterns of pixels that are similar across the data. Then, as performed by Ribeiro et al. (2016), the system can be asked to explain the differences between a dog and a wolf, according to its understanding of the labels. In Fig. 2.4 the system has been asked to explain why a Husky has been identified as a wolf. The explanation points to the pixels that show snow, the environment in which wolves are often portrayed.

If, instead of wishing to categorize and differentiate between dogs and wolves, engineers want to make decisions about people and their actions, where the input concerns information about those people and their actions, they run into other problems with the ingredients. Multiple researchers have written about how values are embedded in technology (Agre and Rotenberg, 1997; de Vries, 2010; Elmer, 2003; Friedman and Nissenbaum, 1996; Hildebrandt and Gutwirth, 2008), how biases and values are represented (Eubanks, 2018; Noble, 2018; Benjamin, 2019), and about the ways in which such technologies can be harmful (O'Neil, 2016). Franke (2022) has tried to structure which biases are possible to detect and alleviate in automated decision-making, using Nozick's (1993) distinction between *first-level bias* and *second-level bias*, as well as between *discrimination* and *arbitrariness*.

Humans need to be able to trust the decisions being made about them, or

(a) Husky classified as wolf (b) Explanation

Figure 11: Raw data and explanation of a bad model's prediction in the "Husky vs Wolf" task.

	Before	After
Trusted the bad model	10 out of 27	3 out of 27
Snow as a potential feature	12 out of 27	25 out of 27

Table 2: "Husky vs Wolf" experiment results.

Figure 2.4: Problems with explaining certain ML models (Ribeiro et al., 2016).

the institutions that make those decisions, and trust in AI might have a higher threshold than for other technologies due to the enhanced comparative abilities of AI (Siau and Wang, 2018). In order to uphold human autonomy people need to be able to give informed consent to such processing, and that in turn requires that people understand something about what the systems are doing and what they can do outside of the intended purposes (Fleischmann and Wallace, 2005). By developing ways to explain neural networks, ML or AI systems likely require more than simply the same technological solutions in XAI, such as a 'holistic, multi-disciplinary, and multi-stakeholder' approach (Rossi, 2018). There might also be a need for more empirical research where humans are exposed to the explanations of XAI (Abdul et al., 2018).

2.4 Regulating transparency

It has now been established that transparency can be beneficial but that there are limitations, and that explaining things is not as easy as it may seem. The technological foundations for why transparency is becoming increasingly relevant have also been covered. As shown in chapter 1, these circumstances have pushed both the public and legislators towards demanding and creating more regulation. This section explores regulation from two different perspectives. First, regulation in a broad sense but specifically applied to policy regarding technology, and

second, the efforts to regulate transparency in law.

Baldwin et al. (1998) argue that regulation can, it its strictest form, be seen as "the promulgation of an authoritative set of rules, accompanied by some mechanism, typically a public agency, for monitoring and promoting compliance with these rules". In a broader sense then, regulation "considers all mechanisms of social control—including unintentional and non-state processes to be forms of regulation." (Baldwin et al., 1998). Both these perspectives will be covered.

The pathetic dot

In the broader sense, regulation can be described as forces that constrain behavior in various ways. One model describing such regulation is the 'pathetic dot theory'.[2] The model deals not only with the law or the social norms that govern society, it also shows how different forces act on an object and regulate it by different means. The theory, published in various iterations (Lessig, 1998, 1999, 2006), had a significant impact on tech policy in general, and tech policy professionals in particular (Allan and Berild Lundblad, 2021a).

The object in question could be anything: Seat belt usage, pirating software, littering in parks, or information transparency. Lessig started by modelling a dot, a pathetic dot, as seen in Figure 2.5.

The pathetic dot is constrained by four different forces, or modalities of regulation. Law constrains by the rules and regulations that governments impose. Norms constrain through the social interactions humans have, and the traditions they create. Markets constrain through supply and demand. Architecture constrains through building either physical or digital restraints of what is possible to do. The main novelty the theory introduces, as has been alluded to, is the argument that code is a form of architecture. Code written to create software and govern how the internet works is a form of architecture that only a select few have the ability to change. Most of us are constrained by how the code has been written and are unable to change it any meaningful way. In these ways, code as architecture is in line with how Winner (1980) thinks about politics in artifacts—namely, that designed objects "embody specific forms of authority and power". This is equally true for code as for physical systems, and therefore it could be argued that code has politics (a political dimension).

Law and norms, according to Lessig (2006), are *ex post* regulators, imposing sanctions on actions or behavior after they occur. For Law, the sanctions are usually fines or jail. For Norms, the sanctions are the menacing glares from a

[2]The name comes from how the model is introduced: "There are many ways to think about 'regulation'. I want to think about it from the perspective of someone who is regulated, or, what is different, constrained. That someone regulated is represented by this (pathetic) dot—a creature (you or me) subject to different regulations that might have the effect of constraining (or as we'll see, enabling) the dot's behavior. By describing the various constraints that might bear on this individual, I hope to show you something about how these constraints function together. Here then is the dot." (Lessig, 2006, p. 121f). Lessig himself calls it 'The New Chicago School', but that name did not stick.

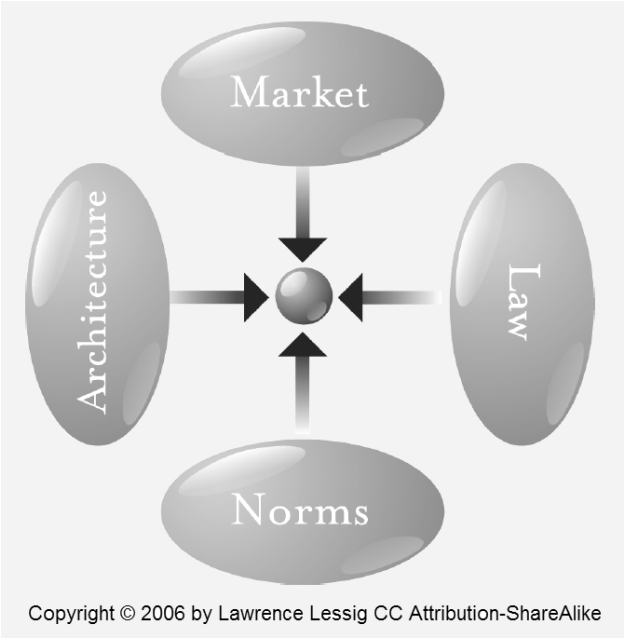

Figure 2.5: A pathetic dot and the modalities of regulation (Lessig, 2006).

neighbor, the ostracization from a social context, or simply the frames of thought that restrict behavior. Architecture and Markets are *ex ante* regulators, regulating before something takes place: architecture by limiting what actions are possible in a given situation, driving through a brick wall will probably not get you to where you are going any faster, even if it is the most direct route. Markets, on the other hand, regulate through pricing mechanisms and supply and demand.

Transparency, as an instance of the pathetic dot, is constrained in all these ways. Law constrains transparency partly through demanding that data processors need to be transparent with certain types of information, and that citizens have the right to such information, and also by regulating that certain information cannot be shared publicly to protect privacy (enabling privacy through regulating information transparency, as Turilli and Floridi (2009) might put it). Legislation covering business practices, financial practices and competition also regulates how information ought to be shared. Norms constrain transparency through the demands that the public has on trust, fair practices, privacy and other values of interest, as well as through the learned practices of designers and engineers building the systems. Markets constrain transparency by the cost of labor to produce explanations for every system, making certain information strategies viable or non-viable. Finally, architecture constrains transparency by structuring what data can be made openly available and how, where especially advanced AI

systems, as discussed in section 2.3, severely limit what a data processor can be transparent with.

The simplicity of the model also means that there are grey areas. The Law limits your behavior due to the fact that you know certain acts will lead to a sanction, and as such acts as *ex ante* regulation as well. The constraints also act on each other, making certain regulatory options more desirable in a given situation compared to others. An example of this can be seen in the GDPR. Some interesting, and difficult to implement, paragraphs in the legislation say that any system that uses personal information must be built according to the principles of *privacy by design* and *privacy by default*. Without going into the exact meaning of those principles, this means that the legislator has decided that the design of digital systems needs to have certain properties. Law has constrained Architecture.

The pathetic dot theory is better imagined as a framing device and a reminder that there are other forces affecting a thing than merely legislation, rather than as a recipe for policy creation.

Legal perspective

As both Baldwin et al. (1998) and Lessig (2006) point out, regulation is often thought of in terms of legal limitations. In regulating technology and transparency, the General Data Protection Regulation (GDPR, 2016) is one of the most impactful pieces of legislation around. The specific legislation is an instance of what Damro (2012) calls *Market power Europe* and what Bradford (2020) calls the *Brussels Effect*, where the EU uses regulation of its valuable internal markets to affect practices in other regions, i.e., the EU uses Law to change Architecture in order to change Markets and Norms. Due to the GDPR being a legislation that enshrines the rights of European citizens, wherever it is being processed, it has significant effects for business in other countries as well.

Many legal scholars have argued about the implications of the legislation since its inception, and due to the relatively short period of time in which it has been implemented and the relative slowness of legal procedures in Europe, there are many aspects of the transparency requirements that have yet to be determined by the courts.

There are now several articles where legal scholars and multi-disciplinary teams including legal scholars have tried to map out the requirements for explainability in the GDPR (Brkan and Bonnet, 2020; Hamon et al., 2021; Bibal et al., 2021). The three papers cited here are all in the cross section between law and computer science. As such, they also share similar features—namely, the same interpretation of the requirements on explanation in the GDPR (which differs from the interpretations in Paper III) even though they end up with somewhat different solutions to the problem.

Hamon et al. (2021) investigate the requirements for explanations according to the GDPR and present alternatives for different audiences, times and expert

levels in the same case, showcasing how different explanations might work in the different scenarios. In a similar vein Brkan and Bonnet (2020) investigate the legal and technical feasibility of explanations in algorithmic decisions and settle on six different types of explanations. Bibal et al. (2021) also analyze the explainability requirements from a legal point of view, concluding that explainability can be applied on four levels: "(i) providing the main features used to make a decision, (ii) providing all the processed features, (iii) providing a comprehensive explanation of the decision and (iv) providing an understandable representation of the whole model."

All three of these, however, lack connections to the softer points of explanations and explainability as mentioned in section 2.2. In that perspective, Wischmeyer (2020) argues that it might be easier than many assume to regulate AI, because the law is already accustomed to dealing with partially opaque and flawed human decision-making, reinforcing the point made by Zerilli et al. (2019) in section 2.2.

Abiteboul and Stoyanovich (2019) aim to reach out to the broader data management community with an analysis of the GDPR, the New York City Automated Decisions System (ADS) Law and the Net Neutrality principle, and discuss how they relate to the data management field. They conclude by stating that "legal norms cannot be incorporated into data-driven systems as an afterthought. Rather, we must think in terms of *responsibility by design*, viewing it as a systems requirement."

Some authors also look at adopting different regulatory environments for AI, such as copying the governance model of the internet (Gasser and Almeida, 2017) and utilizing international human rights legislation to aim for accountability (McGregor et al., 2019). Butcher and Beridze (2019) offer an overview of different governance models and initiatives, and Dafoe (2018) proposes a research agenda for AI governance.

Looking specifically at the paragraphs of the GDPR studied in Paper III and Paper IV, Temme (2017) criticizes the regulation for not addressing the challenges to render algorithms more transparent to a greater extent, and for the lack of clarity about the legal basis of the right to explanation. Wachter et al. (2017) claim that the right to explanation does not exist and should instead be considered a right to be informed, while Selbst and Powles (2017) say that the term should be meaningful information about the logic involved, which is the formulation in Article 15 (1)(h) (GDPR, 2016). On the other hand, Bottis et al. (2019) claim that there is a tension between the right to the protection of personal data and the right of access—similar to de Laat (2018) who argues that privacy is a possible limitation of full transparency.

As a final note, it is interesting to see that Massey et al. (2011) have performed experiments on computer science students in order to discover to what extent they can properly assess whether a certain system aligns with the legal requirements that a computer system has to adhere to. They find that students mostly cannot and additionally point out that the legal experts they consult ensure that the

requirements are well established are not in agreement on what the requirements actually mean—certainly an illuminating insight.

2.5 Designing systems

In section 1.2 definitions of usability and satisfaction were shows, and it was noted that less attention has been given in the research community to the study of satisfaction and how to achieve it (Hassenzahl, 2001; Bevan, 2010), compared to efficiency and effectiveness, the other constituent parts of usability. Additionally, a relationship between satisfaction, on one hand, and explanations on the other, was established by way of Norman (1988) and gulfs of evaluation and execution, which required "bringing into a structural alignment designer intentions, the user interface and user mental models".

This section looks at efforts to bring these things into alignment. To do so, it extends the lessons from sections 2.1 & 2.2 and applies them to the technologies described in section 2.3. It also views these applications through the lens of the regulations in section 2.4. For a broader discussion on how HCI deals with explainable, accountable and intelligible systems, see Abdul et al. (2018).

Building transparency

When describing transparency as a *pro-ethical condition*, Turilli and Floridi (2009) argue that whether the potential user benefits from transparency is contingent on availability of information, accessibility of information and how it can be used to support decision-making processes. The information providers "shape such factors by choosing which information could or should be disclosed, [...] and by deciding in which form information might be most suitably made available" (Turilli and Floridi, 2009, p. 106). Turilli and Floridi also argue that an important element of transparency ought to be transparency with the process of creating information. Without the ability to know how the information transparency has materialized, certain ethical values will not be realized, which is in line with the idea of 'post-interaction' HCI from Comber et al. (2019). Transparency in this regard will likely also have an impact on how well users' expectations and preferences align with how a system functions, if it is used to realize that specific ethical value.

An example of this is shown in Andrus et al. (2021), where the authors interview practitioners working in algorithmic decision-making about the availability and use of demographic information. In the specific context, demographic information can be used to verify whether a data set is biased against or towards certain groups in society. However, such information also has an impact on privacy and fairness which, again, is close to how Comber et al. reasons. Andrus et al. map out several challenges that interviewees face and conclude by saying that one strategy for dealing with bias might not be to increase the availability

of demographic data, but to discuss whether such data should be collected and used in the first place. They argue for changes regarding how the practitioners document and work with data sets, rather than arguing for changes in the technologies.

The same train of thought is taken by Hutchinson et al. (2021), who create a model for data set accountability and argue for the entire ML field to rethink how data sets are constructed, and that data set engineering should be elevated to a key component of AI engineering. They point out the different cultures of data set engineers and the ecological challenges of data set engineering, and they label data sets as infrastructure, much like the pathetic dot theory (Lessig, 1998).

Cysneiros and do Prado Leite (2020) look at non-functional requirements (NFR) in software development with the goal of finding working requirements that can increase trust in automated systems. They identify ethics, safety, security and privacy as NFRs that through transparency can be used to achieve trust, reminiscent of the arguments in Turilli and Floridi (2009). As shown in section 2.1, these results are expanded on in Kwan et al. (2021) where the NFRs' trust, ethics and transparency are mapped out to a fuller extent with interdependent values. Kwan et al. show that transparency has several constituent parts, such as transparency in production, post-disaster transparency, organisational transparency, and procedural transparency. They also identify other values that have positive associations with transparency, such as accountability, usability and clear ethical guidelines. These constituent parts and values all coalesce to build and improve transparency, and in turn improve trust. It is, however, unclear how one should operationalize these concepts.

Barclay and Abramson (2021) investigate the requirements and responsibilities of various roles for an AI app. The app is used by a domain practitioner, and built by a systems integrator; the machine learning model is built by a machine learning engineer, and the data sets are constructed by data scientists. Between these different roles there needs to be transparency in how the different parts have been constructed both down the line and back up the hierarchy. This model clarifies three trust frontiers where each role has to rely on the other roles in order to properly do their job, as well as how information and privacy requirements flow between the roles. Ahmad (2021) presents early progress on a project intended to investigate how requirements engineering can be better aligned with AI development. In a literature review Ahmad has found several aspects, such an absence of communication between software engineers, data scientists and machine learning specialists.

Schneier (2019) looks more closely at the blockchain debate, where several proponents of the technology argue that the blockchain in itself makes the need for trust irrelevant. With the blockchain, there is no longer a need to trust banks, other institutions or other parties. Because the information about every transaction is transparent, users can independently verify all information needed for each transaction. However, as Schneier points out, this hinges on a very specific definition of trust—trust as verification—and such a definition hardly

covers all that is implied by trust. Schneier argues that there are four elements of trust: 1) morals, 2) reputation, 3) institutions, and 4) security systems. What the particular transparency of the blockchain does is "shift some of the trust in people and institutions to trust in technology." (p. 4). Even with the emergent technology a person needs to "trust the cryptography, the protocols, the software, the computers and the network." (p. 4). So using transparency to achieve trust in this case only shifts some of the trust from one part of the trust ecosystem to another.

Fleischmann and Wallace (2005) have been mentioned previously. However, the main purpose of their paper is to rethink how systems should be designed with transparency. The central argument of Fleischmann and Wallace is that along with a covenant with reality when designing AI systems, in that it must be faithful to reality, and a covenant with values, in that models need to be faithful to the values of a client, there also needs to be a covenant with transparency. "The emphasis on transparency not only allows modelers to live up to their ethical obligations, it also undermines the power inequality, since an informed user is in a better position to evaluate a model" (Fleischmann and Wallace, 2005, p. 97).

Building explainably

Looking instead at how explainable systems are built, or how to build better explanations, Langer et al. (2021a) argues for a multi-disciplinary approach to auditing explainability. If explanations are based only on the experiences and understanding of a single discipline, say computer science, then an auditing of those systems will be lacking. Specifically, they argue that the multi-disciplinarity should consist of technical, psychological, ethical and legal perspectives.

In another paper, Langer et al. (2021b) argue that the main goal of explainability is to satisfy certain wishes that stakeholders have on an artificial system. These *stakeholder desiderata* have, however, not previously been properly mapped, and by providing such a mapping Langer et al. (2021b) hope to improve the intra-disciplinary understanding in XAI. Chazette et al. (2021) investigate explainability in requirements engineering and develop both a model for it and a knowledge catalogue showing how explainability affects other quality attributes.

Binns et al. (2018) investigate design of explanations, specifically for the requirements of "meaningful information about the logic involved" in automated decision-making as stipulated in the GDPR—a requirement that is explored in Paper III and IV in this thesis. They design four styles of explanation: *Input influence-based*, *demographic-based*, *case-based*, and *sensitivity-based*. Binns et al. (2018) then test these explanations on users in order to find to what extent users perceive them to be just. They find that sensitivity and input influence-based explanations are perceived as most just. The sensitivity-based explanations especially resonate with users in interviews due to the perception that it provides information on how to change behavior in order to get different results, a perspective that validates de Laat's theories (2018). Binns et al. has been a sig-

nificant influence on how to consider the design of explanations throughout the thesis project. In a similar fashion to Binns et al., Sadeghi et al. (2021) create a taxonomy regarding different points of failure in a system, and what types of explanations are required in each scenario.

Alvarado and Waern (2018) introduce the concept of Algorithmic Experience (AX), trying to make explicit the interaction with and experience of algorithms. Regarding the requirements for AX-design, they create five functional requirements: profiling transparency and management, algorithmic awareness and control, and selective algorithmic memory. They argue that AX can make relationships with the service "more joyful", likely raising satisfaction and thereby improving usability (ISO, 2018). Ehsan et al. (2021) create an explanation model for AI systems where the explanations given to a user are based on previous users' interactions with the system. Cheng et al. (2019) show that interactive explanations and "white-box" explanations (which, like Turilli and Floridi (2009) argue, show the inner workings of an algorithm) improve users comprehension of a decision.

In trying to improve the transparency of news algorithms, Rader et al. (2018) design and test various explanations of the Facebook News Feed algorithm. They found that while explanations increased understanding of how the algorithms worked, it was difficult to create explanations that helped users evaluate correctness or form opinions on whether the algorithms were sensible and consistent.

2.6 The theoretical contribution

This chapter began by introducing Turilli and Floridis 2009 model for transparency as a *pro-ethical condition*. Operations on data creates information, which in turn enables or impairs ethical values in different ways. This process was presented in Figures 2.1 and 2.2.

Section 2.2 looked instead at how explanations work, specifically covering the theory of explanations presented by Keil (2006). This is an important puzzle piece in how transparency creates the values different actors wish to achieve. Explanations help humans increase understanding of a phenomenon (Wilson and Keil, 1998). Information transparency therefore needs an element of explanations in order to facilitate the decision-making process. The decision-making, in turn, needs to adopt or design explanations in order for the recipient to understand the information—and for the ethical value to be realized. This model of transparency was shown in Figure 2.3.

In short, while information transparency is a pro-ethical condition that enables or impairs ethical values, designers or businesses need to apply explanations to that information to help them make decisions regarding the information. They also need to design explanations for other actors and consumers in order for the ethical values to be achieved.

The subsequent sections in this chapter present the contexts, and extant research, in which the transparency and explanation model is applied, namely the algorithms and artificial intelligence that underlie the automated decisions with which data controllers are supposed to be transparent. The regulations, both general and legislative, govern what data processors must be transparent with. Finally, the design processes by which the systems and transparency is created are expounded.

Chapter 3

Methods

As mentioned adove, the constituent parts of this thesis are built on different methods for different types of papers. In each case, the authors first came up with a problem description to investigate, and select an appropriate method to investigate the problem.

3.1 Text analysis of government strategies

Paper I studied to what extent companies and other actors seeking to adopt AI-technologies, or other automated decision-making systems, received guidance from government strategies and policy documents concerning AI, specifically on ethical and responsible AI. The Nordic Council adopted the position that the Nordic countries have a competitive advantage in making AI more ethical, and the paper therefore investigated how the Nordic countries tried to realize this ambition in various strategic documents.

First, the authors identified the strategies produced, and which countries were possible to study within the given context. The Nordic Council consists of Denmark, Finland, Iceland, Norway and Sweden, the autonomous territories of the Farœ Islands and Greenland, and the autonomous region of Åland. The authors are fluent in Swedish and English and understand written Danish and Norwegian to an acceptable degree. The Danish, Norwegian and Swedish documents were produced in their respective languages, but the Finnish documents were produced in both Finnish and English. There were no relevant strategies from Iceland written in English, nor, at the time, in Icelandic, nor any strategies from the territories or Åland. As such, the final selection came down to Denmark, Finland, Norway and Sweden. For each of the countries selected, two strategies were chosen for the study. Since the policy area was and is fast moving, the selection had to change during the process, as newer and more relevant reports were published in Denmark, Norway and Finland.

As a framework for analysis, the AI4People framework was adopted (Floridi

et al., 2018). The AI4People principles are as follows: (i) beneficence, (ii) non-maleficence, (iii) justice, (iv) autonomy and (v) explicability. The authors read the documents in turn, reading half the documents each. For each idea, concept or phrase in the texts that fit into one or more of the ethical principles, the authors marked the phrase. The authors then read the documents that had been read by the other author and either marked new findings, questioned marks made by the other author, or added marks for concepts that could fit into other principles. After this, all marks were combined in a separate document and grouped in different classifications (Esaiasson et al., 2007, p. 238). The authors met throughout the process and discussed how the groupings were made, as well as evaluated and re-evaluated the marks, until a consensus on how to interpret different concepts and what categories each of the concepts belonged to was reached. This made the interpretations much more grounded than they would have been with a cursory glance. The categories were then presented in the graphs, which can be seen in Paper I.

3.2 Data collection through consumer requests

In Dexe et al. (2020), which Paper III and IV are partly based on, the authors developed a new method to collect information about transparency in practice. It attempts to empirically test how the GDPR was being adopted, and how companies interpreted the transparency requirements. At the time no other such methods had been published. However, simultaneously, Sørum and Presthus (2020) developed a very similar method which informed a more rigorous application of the method in Papers III and IV. There is also a late breaking work by Krebs et al. (2019) that seems to be utilizing similar methods but has yet to produce a final result. This is a novel form of data collection, and one that could be applied in other efforts to investigate transparency claims. Others have approached the GDPR empirically (Alizadeh et al., 2020; Bahşi et al., 2019; Machuletz and Böhme, 2020; Sanchez-Rola et al., 2019; Nouwens et al., 2020; Momen et al., 2019; Fan et al., 2020; Syrmoudis et al., 2021), but none using this methodology.

Original study

Dexe et al. (2020) aimed to find a way to collect data about transparency practices that was not reliant on the willingness of companies to co-operate with research projects. This meant extensive negotiations regarding access to development teams and respondents within the companies. One avenue that seemed promising was to utilize the requirements in the GDPR that companies had to provide access to information about automated decision-making processes. Right to access requests made by real consumers could give real world practices concerning transparency. There are other ways the requests could have been made, which is acknowledged in Dexe et al. (2020) and Paper III. There are ethical questions to consider here, the main being whether it is ethical to use information for

research that companies have given to consumers without an expectation that it could be used for research. First, there are no stipulations in the GDPR about what the rights holder is allowed to use their information for. Second, it is reasonably the right of the consumer to share explanations given to her in whatever way they choose.

Volunteers were recruited and sent requests to the companies in whatever way they thought most appropriate. For some this meant filling in online forms or sending a customer service email, while for others it was chat boxes. The researchers then collected the responses and, analyzed what details were in the responses, as well as the points concerning how the responses were written. These were then summarized as categories by which different responses were compared. These tables can be seen in Dexe et al. (2020) and in Paper III.

In order to be able to reproduce the study in Paper III, the method was summarized as follows (Dexe et al., 2022):

1. Identify the relevant type of insurance—extant insurance offerings will differ between countries and companies.
2. Get an overview of the market—identify main actors and rough market shares (to know which ones are most relevant to include).
3. Translate the request.
4. Recruit volunteers—making sure that there is only one per company.
5. Send requests—volunteers are asked to note the date and means of contacting their insurer.
6. Gather responses from the volunteers and analyze these.

Sørum & Presthus's approach

While research for Dexe et al. (2020) was being conducted (between December 2018 and March of 2019, with the article being submitted in March 2020), Sørum and Presthus published a very similar study on the Norwegian consumer market (the article was submitted in October 2019). While Dexe et al. (2020) focused on the insurance market, and only on art 15 (1)(h) of the GPDR, Sørum and Presthus (2020) requested information on both the full extent of art 15 in the GDPR and art 20 (the right to portability). They directed requests to 15 companies across different segments of the Norwegian consumer market, with varying size, domain and nationality. As with (Dexe et al., 2020), they chose companies with which they had an ongoing customer relationship, in order for the requests to be valid under the GDPR. Article 15, right to access, is described in detail in Paper III and IV. Article 20 gives the consumer the ability to move their information from one service to other, competing services. This requires that the information be made machine readable in order to be transferred correctly. Sørum and Presthus

also chose a formal but easily understood request, although they did not aim it at a specific type of processing. As with Dexe et al. (2020), Sørum and Presthus did not send reminders or clarifications to the companies.

It is noteworthy that only three of the companies approached by Sørum and Presthus gave explanations in relation to art 15 (1)(h). Accompanying the article is a list of recommendations for consumers who wish to request access to their information, including advice for consumers to 'clearly state' what type of information they want access to.

The Sørum and Presthus study can be characterized as a *wide range* study (general requests to a wide range of companies), and Dexe et al. (2020) as a *limited range* (narrow requests to a single sector) study.

Adaptations in Paper III

Paper III is a direct continuation from Dexe et al. (2020). Through contacts, five additional researchers were contacted to conduct replication studies in four additional countries: Denmark, Finland, Poland and the Netherlands.

As shown, the steps of the original study were summarized to make replication easier. I also had dedicated meetings for each country to discuss specific adaptations necessary and clarify different aspects (such as what is actually included in 'home insurance' in the different markets).

However, few compromises were needed in the adaptation of the method. The most common change was that the researchers used other types of networks than their work place. In Poland, a few volunteers were recruited through a think tank, but otherwise the recruitment was done through personal contacts in some way. In each case the researchers either had legal training or consulted legal expertise in their respective countries in order for the translation of the request to be as accurate as possible in terms of legal coverage.

The results from Dexe et al. (2020) were re-used as the Swedish sample for Paper III. The only new information presented Paper III that related to the original study was an example reply, which came from one of the authors participating in both studies—all other data in Paper III were taken from the already published results in Dexe et al. (2020).

Adaptations in Paper IV

If Paper III and Dexe et al. (2020) used narrow requests on a limited range, and Sørum and Presthus (2020) used general requests to a wide range of companies, Paper IV combines the perspectives by making a narrow request to a wide range of companies. The paper uses a request similar to Paper III, but with necessary alterations, and uses a similar scope to Sørum and Presthus by looking at 24 different companies in a number of different business areas. Another similarity with the Norwegian study is that, because of the broad range of companies, Paper IV only includes companies where the authors have a customer relationship.

Some adaptations were made in comparison with Sørum and Presthus (2020). First, requests were ideally only sent to companies that had an office in Sweden. Second, the sample was limited to companies assumed to have some kind of automated decision-making process in place. The reason for not actively investigating whether such processing actually happened was that consumers will likely send requests based on their assumptions rather than after a thorough investigation into the various technological processes the companies employ. The authors also decided to send reminders to companies that failed to reply and engaged with those responses that were unsatisfactory or where the wrong information was given. As such, for Paper IV several clarifications were sent to companies.

3.3 Interview-based studies

Paper II is concerned with the views and opinions of insurance professionals and how to gauge the impact transparency has and can have on their business. As such, the recruitment of respondents consisted of the authors contacting insurance companies and asking for interviews with various domain experts. The end result was 8 experts from 4 companies, in middle or upper management, either in product development or in consumer relations.

For Paper IV a number of reference interviews were conducted, three with companies that data requests had been made to, and two with other organization (one consumer protection authority and one trade association).

Interview methodology

For both papers, the approach to the interviews themselves was identical. Both aimed to have respondents describe how they think about the context in which they are in (Esaiasson et al., 2007, p. 285). The authors used prompts presented to the respondents during the interview. The prompt in Paper II was presented in the beginning of the interview and consisted of a working definition of transparency, as well as results from an opinion poll about demands for more transparent services. The prompt in Paper IV was presented near the end of the interview and consisted of examples of different styles of explanations, adopted from Binns et al. (2018).

Both studies used a semi-structured approach to interviews. For both, a document with a number of main questions or themes was prepared, with a number of questions under each main question (Esaiasson et al., 2007, p. 298). The template for Paper II is available in the paper, and for Paper IV only the overarching themes are included.

In all interviews (except one, where a co-author did the interview on their own) the first author was the lead interviewer. The semi-structured approach leaves room for interrogations of noteworthy replies to previous questions, and makes it possible to discover lines of questioning that were not or could not

be prepared before the interviews. It also means that the unexpected lines of reasoning from previous interviews inform and affect the lines of reasoning in future interviews, even though the themes and headlines of the interview stay the same. The interviews were all recorded and then transcribed for further analysis.

3.4 Opinion polls

Finally, in Paper II, and also as input for the whole research project, several polls have been used. For Paper II, one poll was conducted specifically for the paper, and the other was published by Insight Intelligence in which I have participated as an expert since 2015.

The Delade Meningar polls are set up by the insights agency Insight Intelligence. Each report is funded by a consortium of four stakeholders, which also means that most of the questions are different in each report. I have participated as an independent expert in all reports since the first in 2015. The poll is administered by SIFO Kantar, one of Sweden's largest polling companies, who polls 1000 people in their online panel. The sample is representative of Sweden as a whole. The other poll was done in conjunction with the insurance company Länsförsäkringar as 200 of their employees took the Elements of AI course. Due to time constraints and employee workload limitations, a single question was asked to all the employees who undertook the course, at the time they finalized it. The poll gives a general indication about the perception of insurance professionals but is only used as a general input in Paper II due to the limited scope and depth of the poll.

Chapter 4

Results

Paper I—Nordic Lights? National AI policies for doing well by doing good

In 2018 the Nordic Council of Ministers for Digitalisation declared the following: "Countries that are successful in utilising and realising the benefits of AI, while managing risks responsibly, will have advantages in international competition and in developing more efficient and relevant public sector activities". With artificial intelligence being at the forefront of almost any tech-policy discussion the statement by the Nordic Council is enticing. Is there really a competitive advantage in fostering more ethical AI applications? And if there is, what are the Nordic countries, many of them known for their technical proficiency, doing to realize this competitive advantage?

These were the questions investigated in Paper I (Dexe and Franke, 2020). By analyzing AI strategies in Denmark, Finland, Norway and Sweden through the lens of the Floridi et al.'s AI4People principles (2018), the authors studied how Nordic governments considered the relation between AI principles and competitive advantages. The research questions concern which ethical values are reflected in the strategies, and how the links between those values and a competitive advantage are described, including what concrete measures are proposed to realize that advantage.

In Paper I there are both detailed tables of all the findings and an extensive digital supplement listing every single mark made in the documents. The tables show that there are plenty of instances in which different ethical values are reflected in the texts. Most prominently beneficence, doing good, plays a large role. AI will be doing good by promoting economic growth, innovation, more efficient services in general and better public services in particular. The dark clouds are accordingly less prominent—non-maleficence, not doing harm, mainly takes the form of reiterating that cyber-security is necessary, and warning about malicious use or misuse of AI. Autonomy, the individual's ability to act on their own volition without outside influence, is largely absent in the Swedish documents despite

being very relevant in the other documents. Denmark, Finland and Norway all mention the importance of data ownership, AI literacy among the population, letting humans shape technology rather than the other way around, and the importance of informed choices. Justice, individuals being treated fairly, is a mixed bag, evident from the extensive list in the digital supplement but with just a few marks for each kind of statement. Most prominent are the effects on the labor market and the importance of avoiding discrimination and bias. Finally, explicability, the ability to explain and be transparent with technology, is the new ethical principle introduced by Floridi et al. (2018). Here are discussions about accountability for AI technologies, the use of open public data and a regulatory environment that is clear and easy to comply with. Finland, Norway and Sweden all mentioned transparency specifically, while Denmark received marks for associated concepts such as the ability to audit and trust in AI. Finland and Norway also received marks for those, unlike Sweden.

Unfortunately, despite the ethical values being present in the AI strategies, they do not give the reader—be it a company or a concerned citizen—guidance in how to actually make sure that the AI is ethical, and that these ethical values are reflected in the products that will eventually be created and applied. In almost all cases they are present only implicitly—the statements made reflect an ethical value, they do not propose specific ethical values openly.

On the other hand, that was probably never what the documents were intended to do. That is, they were not created to give concrete guidance to practitioners. Instead, they were likely meant to affect other policy makers. For the advisory bodies, this seems fairly straight forward. They are meant to create documents that contain advice that policy makers listen to. For the government bodies, mostly ministries of innovation, business or financial affairs, the documents lay the foundations on which more concrete policy development is later built. Only the Finnish documents contain concrete suggestions written and endorsed by a government body, and even those mainly concern various types of funding or further investigations or reports.

As for the links and measures, the authors initially assumed that if the documents contain some link between ethical values and a competitive advantage, then by default, there would also be a concrete measure described to enable that link.

This proved not to be the case.

The links that were most prominently featured were the first-mover advantage, claiming to be the first in applying AI, and making sure that those applications are also ethical, which means that others will apply the same, ethical, models simply due to their being available. This was the only link present in documents from all countries. Other possible links include being able to balance the benefits and risks of AI in a responsible manner, or making sure that the digital economy is also built through ethically sustainable practice. In fact, the Danish strategies propose *making* ethics a competitive advantage in and of itself.

The measures described to enable such links are, as already mentioned, not

actually linked to the links. The most promising measure is thus for Nordic countries to influence international standards and establish a market demand for ethical solutions (which could be linked to the link above, but they appear in different documents). Test-beds and regulatory innovation also feature in the documents, as well as the ability to affect the market through public procurement processes.

It is highly unlikely that the AI strategies analyzed in Paper I will mean that the Nordic countries will realize the competitive advantage of applying ethical AI in their economies. That is not to say that such a competitive advantage does not exist, or that the Nordic countries will not be the ones who do manage to achieve it. It is just that these documents will not be the guiding light making sure it happens.

Paper II—Transparency and insurance professionals: a study of Swedish insurance practice attitudes and future development

In Paper II (Dexe et al., 2021) the authors wanted to understand what companies and the people working in them think about the possibilities of utilizing transparency as a competitive advantage. More precisely, what do insurance professionals think about the possible benefits or drawbacks of transparency in insurance?

The problem is based on a question from Delade meningar (2019) seen in Figure 1.1 in section 1.1. When asked about the importance of transparency as a deciding factor when choosing new services, 63% of Swedes say that a more transparent alternative would make them consider switching services, and 47% of those would consider this even if the new service is more expensive. While these numbers should not be taken at face value, they do point to a desire that insurance companies (and other companies and organizations) might use in order to gain advantages over their competitors.

More precisely, Paper II asks how the insurance companies view the competitive advantage of transparency, whether they use transparency as a strategic tool, which limits the transparency they can identify, and to what extent AI plays a role in their business.

The data in Figure 1.1 was used as a prompt, together with a working definition of transparency that can be seen in Paper II, in the eight interviews conducted in the study. Together with a custom poll described in section 3.4, these give insight into what value transparency has and could have for the insurance industry going forward.

Of the 200 insurance professionals polled in the custom poll mentioned above, 74% of the respondents said that "transparency and openness in AI decision-making can be a competitive advantage" for their company. Do the insurance professionals agree?

Sure. All respondents thought that transparency is beneficial, and most of them think it can be a competitive advantage. However, the respondents have

very different perspectives on what "transparency and openness" means and which parts of the insurance business should be transparent. Some respondents claim that their respective companies are transparent because they provide customers with all the necessary information about the services before a customer signs a contract (for instance, the terms of service documents and other legal documents). Others claim that their company is transparent because they manage the information and try to explain terms and conditions in plain language rather than giving the customer the full legal documentation directly. The terms and conditions are therefore both transparent and opaque, and explaining them in plain language could be more transparent than showing the full text, depending on who you ask.

Transparency plays a significant role in *trust building* according to the respondents, and several of them mention the dictum "insurance is in the business of trust". Most also agree that one of the most important applications of transparency is in *expectation management*, meaning that transparency can help make sure that expectations are not violated, causing discomfort and mistrust (Kang and Hustvedt, 2014). One respondent argues that even though transparency may have these effects, for consumers it is probably only important on the margins—in most cases the price and value of the service will play a bigger role in the choice of the consumer than the existence or lack of transparency. The respondent adds that while there might well be some groups of potential customers that see it as a selling point, echoing Bhaduri and Ha-Brookshire (2011), they are likely not a significant market segment.

Several respondents discussed that for some types of insurance there is almost no drawback to being transparent. This mostly concerns insurance products that are very similar across the industry, such as home insurance, but not when it comes to price. Information about pricing (not the price itself, but how it is set) was the limitation of transparency most commonly referred to across the interviews. The reason given is that since most of the insurance products are very similar across the industry, thus pricing is the main point of competition. Disclosing the pricing algorithms would make it possible for competitors to under cut the price point of a particular insurance.

As for the question of AI and insurance, the authors were somewhat surprised by the findings. The entire starting point of the research project in which Paper II was conceived was investigations of transparency in relation to algorithms and AI in insurance. However, almost none of the respondents seemed comfortable discussing how transparency will be affected by the use of AI in insurance, for instance by using black box models. This suggests either that the use of AI in insurance is not integrated with the product development and customer relations side of the business, where the experts where situated, or that insurance companies are not yet using such advanced artificially intelligent systems as debates and news articles sometimes assume. The latter scenario seems to be the most likely.

In the end, there was no agreement found on what advantage transparency

would have in the insurance industry, nor any evidence of the strategic use of transparency. The respondents saw benefits, sure, but also several limitations, and were not as close to adopting more advanced AI technologies as suspected.

Paper III—Explaining automated decision-making—A multinational study of the GDPR right to meaningful information

Paper III (Dexe et al., 2022) investigates how insurance companies work with transparency in practice. Specifically, it examines how they work with legally mandated transparency in the GDPR, and the requirement to provide "meaningful information about the logic involved" in automated decision-making systems. While the companies are unsure of how to realize a competitive advantage with transparency, they still must be transparent with various processes.

Paper III is based on Dexe et al. (2020), as mentioned in section 3.2. The original paper investigated the practice of the Swedish insurance market in responding to requests for access to information according to article 15 (1)(h), specifically regarding how pricing (which, remember, respondents in Paper II were reluctant to be transparent with) in home insurance is done. Paper III expanded the study to include four reproductions of the original study in four additional countries: Denmark, Finland, Poland and the Netherlands.

Paper III extends the results from Dexe et al. (2020) in three important ways: (i) the increased scope adds to the generalizability of the results, (ii) the comparison of different language versions highlighted problems with interpreting the relevant articles in the GDPR, and (iii) a theoretical discussion was added about how to view the quality of the responses.

The expansion of the study to four additional countries was less complicated than assumed. While there might be qualitative differences between the general coverage of home insurance offerings in the different countries, and what type of homes are insured, decent samples were achieved in all countries, the lowest coverage being 40-45% of the market in Poland and 45% in the Netherlands. However, as can be seen in the comparison tables in Paper III, there were no stark differences between the countries. There was no indication of difference in how companies replied in terms of the size of the company, or age of the company or ownership structure. The Danish and Swedish companies were the only ones that provided any sort of logic about how the pricing mechanism worked, but even those explanations were very limited (revealing that one type of data, e.g. address, might mean that the price goes up or down).

While comparing the language versions[1] of the GDPR in the five countries, as well as the English version, it became apparent that there were different possible interpretations regarding what the requirements in article 15 (1)(h) actually en-

[1]No single language in the European Union has legal standing over any other language, so all language versions are equally valid in terms of enforcement.

tailed. In the Dexe et al. (2020) it was simply assumed that the right to access to meaningful information included all types of automated decision-making involving data regarding the individual rights-holder. Specifically, the words "at least in those cases" when referring to "profiling, referred to in Article 22(1) and (4)" had been read as an *example* of what data controllers could be transparent with. In Paper III called this a *broad interpretation* of the legal text. The alternative view, a *narrow interpretation*, instead suggests that the reference to Article 22 contains the full extent of what data controllers must be transparent with, for all other types of processing they would only need to adhere to the principle of transparency in article 5 (GDPR, 2016). Such an interpretation would mean that only "profiling", which has legal or similarly significant effects for the individual would be covered by the right to access in article 15 (1)(h).

Until there is a decision from the European Court of Justice, it will probably remain unclear which interpretation is true, even if most companies adopt a narrow interpretation.

Paper III also looks at available literature regarding how explanations could be improved, and to what extent understanding, trust or sense of justice can be increased by use of explanations. Literature that tries to design explanations that would be compliant with the GDPR (Binns et al., 2018) tends to be too idealistic in their designs. This is not to say that the explanations that they have proposed are bad, but rather that they go far beyond what real world explanations data controllers have settled on. A contrasting view is offered by Wilson and Keil (1998), as described in section 2.2. Applying Wilson and Keil to the responses given by the insurance companies puts them in a different light. The people who have requested information likely know that it is a response about insurance, that they are insurance customers and likely know at least something about what an insurance company does. That is not to say that the explanations are legally compliant, but they might be still offer more insights in their native context than they appear to if read very critically.

Paper IV—Transparency hurdles—investigating explanations of automated decision-making in practice

The idea for Paper IV was to recreate Sørum and Presthus (2020) on the Swedish market, with 24 companies across different industries, but narrow down the scope of the request in the same way as in Dexe et al. (2020) and in Paper III to include only requests for information about automated decision-making. In addition, five reference interviews were conducted.

Out of the 24 companies to which requests were sent, 12 interpreted the request as a generic access to information request, similar to the requests sent by Sørum and Presthus. This meant that rather than giving information about automated decision-making according to Article 15 (1)(h) of the GDPR, they offered us a copy of the personal information they processed about us. Four compa-

nies failed to even give a response, meaning that 60% of the companies that did respond misinterpreted the request.

Eventually, after clarifications were sent out to some companies that misinterpreted the request, 12 additional responses were received. Two companies, after clarifications had been made, gave (lackluster) explanations of their automated decision-making. Four said that they had no automated decision-making, and six stated that their processes were not covered by the requirements in Art 22 of the GDPR—profiling that has legal effects or similarly significant effects on the individual—meaning that they seem to subscribe to the narrow interpretation of Art 15 (1)(h) discussed in Paper III. It is also possible that the four companies that claimed they had no automated decision-making subscribe to the same view, but simply did not refer to it in their response. The differences between Paper III and Paper IV are striking in this regard.

The data collection phase for Paper IV left the authors with a question: Why is it so hard for consumers to gain access to good explanations of how their data is used for automated decision-making?

To begin answering that question, 9 different transparency hurdles were proposed, that could help illuminate why it becomes difficult to gain access to information for consumers. These hurdles are based on experiences from both Paper III and IV, relevant literature such as Sørum and Presthus (2020) and other available literature on the topic:

1. Interaction strategies—Companies need to settle on a strategy to deal with consumer requests. Depending on what strategy they use, specialized requests might be harder to make, as more generic (and therefore more common) requests take priority in the work with designing responses.

2. Targeted language—The specificity of the request in Paper III made it possible for a customer service representative to immediately realize what domain the request dealt with: "home insurance", "premium", and "pricing" are insurance related terms. The request in Paper IV also required such a representative to try to interpret what was meant by automated decision-making and evaluate the companies products on their own.

3. Avoidance—The requirements for transparency in the GDPR might lead to companies simply avoiding automated decision-making, which was acknowledged in the interviews, or process the information in ways such that it becomes easier to claim it does not deal with personal information.

4. Legal interpretations—As described in Paper III there is an interpretational problem with the legislation. Before the European Court of Justice has been able to hear a case that could settle the question of interpretation, customers will likely have to face the fact that companies will probably lean towards a narrow interpretation of the law, and refuse to give information to processes that do not reach the requirements of art 22.

5. Value chain transparency—For a large number of companies on the digital marketplace, the use of automated decision-making will likely come down to whether they purchase such products from other companies. In such cases, the transparency the company can give to their consumers will, in part, be contingent on the transparency between the company and their supplier of technical analysis or tools.

6. Specificity—Tied to the *targeted language* hurdle comes the problem of specificity. Even if requests were sure to not only cite a specific article in the GDPR, but even a specific paragraph within that article, companies seemed to fail to understand the request. Customers might need to be prepared to not only specify the legal basis for their request, but also to explain in detail what that actually entails.

7. Ability to question—Considering the problem with *legal interpretations* and the fact that many companies seem to adopt a narrow interpretation, it becomes difficult for a consumer to question the responses that companies give if those responses are unsatisfactory. The consumer cannot exercise their right according to the broad interpretation if the companies do not acknowledge that right.

8. Expertise requirements—If the consumer has a different interpretation than the company, or if responses are unsatisfactory in some other way, the consumer is suddenly facing a huge hurdle in terms of the expertise required to legitimize other interpretations. Say that there really is a legal effect, or that a system really is fully automated, or that the consumers think that a specific type of data is being used that requires technical know-how and thorough investigations into company terms of service. Is that a reasonable burden to put on consumers?

9. Bad explanations—Finally, a response to the request has arrived, but when reading it, it just comes across as... bad. Sørum and Presthus (2020) seem to have had this problem as well, and even the subjects of various experimental studies have shown a lack of satisfaction with different designs of explanations. If the thoroughly designed explanations do not manage to increase the understanding and trust to the extent the researchers want, and since real world explanations are much less complete than those, then customers might never receive satisfactory responses to requests such as these.

Paper IV argues that practices regarding transparency about automated decision-making fall short of the expectations of consumers and experimental studies. It makes it difficult for consumers to maintain their autonomy and engage in the digital sphere with informed consent. The hurdles point to a number of problems, but are by no means an exhaustive list of why it is difficult to craft and gain access to decent explanations.

Chapter 5

Discussion

In section 1.3 three main research questions for this thesis were presented: 1) what benefits does information transparency generate, 2) why is it so hard to achieve transparency in automated decision-making, and 3) in what ways does transparency relate to usability.

In previous chapters theories related to how transparency and explanations work, what constitutes an algorithm and AI-systems, and how regulation works in both general and specific terms have been presented. In addition, the possibilities of designing for transparency and explainability have been investigated.

Then, the methods and results of the four papers included in the thesis were presented. Paper I investigated what guidance Nordic AI strategies give in ethical questions related to AI, and Paper II asked insurance professionals what they think about the possibilities and risks of information transparency in insurance. Paper III and IV investigated the practice of being transparent, specifically by explaining the logic behind automated decision-making. A number of suggestions for transparency hurdles that might need further consideration have also been presented.

This chapter combines all these perspectives and discusses what has actually been discovered, uncovered or missed.

5.1 What transparency can accomplish

Transparency can be portrayed in a myriad of ways, as seen in section 2.1. Business and information management research often use transparency to mean "the process of making explicitly and openly available (disclosing) some information that can then be exploited by potential users for their decision-making process" (Turilli and Floridi, 2009, p. 105f). Ball (2009) introduced three metaphors by which transparency can be understood: the first as a value that counters corruption, the second as open decision-making and the third as a tool for good

governance. Larsson and Heintz (2020) argue that transparency in AI is associated with trust and accountability, much like Turilli and Floridi.

It seems, in the literature, that transparency has several important effects. The previously mentioned antidote to corruption is an important one, as is the ability to investigate financial wrongdoing (Bauhr and Grimes, 2014; Hood and Heald, 2006). Important as these effects are, they have only a limited application to the context of algorithms, automated decision-making and what extent consumers understand how personal data is being used.

As shown, some ethical principles are dependent on transparency in order to be realized, transparency being a *pro-ethical condition* (Turilli and Floridi, 2009). Accountability, informed consent, and safety are all contingent on information being made available about actions, conditions and risks. Transparency can increase trust, or at least affect it to some degree (Kim and Lee, 2012; Kang and Hustvedt, 2014; Kim and Kim, 2017; Bhaduri and Ha-Brookshire, 2011; Cambier and Poncin, 2020). Kwan et al. (2021) showed a number of goals that are both positively and negatively associated with transparency. In Paper II all respondents mentioned transparency as something desirable and beneficial, at least to some degree, and in Paper I transparency or associated concepts featured in all strategies.

However, there has been little evidence or consensus regarding effects of transparency. There is also little evidence that companies have clear transparency strategies—neither from the interviews in Paper II, nor from the responses in Papers III and IV. If transparency is the intrinsic good it is often portrayed as, and if it was easy to realize the benefits of transparency, surely there would be fewer calls for more transparency either from the public or from legislators. It is therefore necessary to identify what stands in the way of increased transparency.

5.2 What stands in the way?

One conclusion of the research papers included in this thesis, as well as the literature covered in section 2, is that there are limitations to transparency that need to be considered. This section looks at three overarching themes limiting the application of transparency.

Transparency might not always be the right choice

First, there are instances where transparency is the wrong choice. After all, it is not desirable for someone to have access to all information about everything, all the time, and there might be good reasons for that.

In constructing transparency as a *pro-ethical condition* Turilli and Floridi (2009) suggest that transparency has two main relationships with other ethical values. For some, the ethical values are *dependent* on information transparency, and for others information transparency needs to be *regulated*. Both these relations can either impair or enable ethical values. Privacy, to an extent, requires

that some information is *not* disclosed, as does anonymity. Copyright sets limitations on how intellectual properties can be shared. Security requires that certain information is held within certain confines. For all these, more transparency would likely impair the underlying ethical value.

Based on the results, business confidentiality and competition are values that are likely impaired by transparency. In Paper II, several respondents argue against being transparent with pricing algorithms, as well as certain insurance models. For pricing algorithms, transparency would mean that competitors would be able to undercut the price point of the insurance product, and for the insurance model it would mean both that customers could 'game' the algorithm, and that competitors could benchmark against it and create more competitive products. It might be important to note here that transparency in pricing might not be beneficial for customers either. Indeed, too much transparency or the wrong kind of transparency might hurt price competition (OECD, 2017). Paper III showed that some insurance companies referred to business confidentiality as a reason for not disclosing more information in their replies.

Paper IV identified a hurdle labeled *value chain transparency*, which is related to the problem of business confidentiality. Many companies that employ automated decision-making will not create the machine learning models or algorithms in-house, which Barclay and Abramson (2021) also shows. Instead, they will likely purchase those algorithms from another, more specialized, company. Barclay and Abramson showed that this creates several different trust frontiers, where different roles across the value chain need to place trust in other roles, and transparency needs to flow in different directions. However, in a business transaction, there will likely be things that the company developing the machine learning algorithm will not want to disclose to the company buying the product, due to business confidentiality. The company buying the product will therefore be unable to disclose information about those specific processes to their customers in turn.

Lessig (2009) also argued that transparency can have negative consequences. When, say, a politician has to be transparent about who donates financial contributions to their campaign, there is a risk that the information is interpreted in ways that do not give an accurate view of reality. Allan and Berild Lundblad (2021b) expand on this idea by saying that one reason that companies are reluctant to be transparent with certain information is that if there are things that are absent, such as reasons for why a certain action has been taken, the public will tend to interpret the 'hole' in the information in a way that is often worse than the reality of the system, much like in Lessig's example. Lakkaraju and Bastani (2020) and Kwan et al. (2021) discuss a similar problem with skewed explanations that create negative impacts. Lakkaraju and Bastani show that some explainable AI models create explanations that do not reflect what biases are in the system, and as such explanations can lead to unwarranted trust in the system. Moreover, Kwan et al. shows that vague and false transparency can negatively impact trust in a system. False transparency denotes when someone only presents positive

information, without disclosing negative information. While they do not explain what vague information is, it is likely related or similar to fuzzy information (Asadabadi et al., 2020).

Technology makes transparency more difficult

The second reason for why transparency is hard to achieve is that the technologies behind the algorithms are sometimes opaque, sometimes incomprehensible, and sometimes it is difficult to choose how to be transparent with the algorithms.

Section 2.3 showed that the use of advanced machine learning algorithms, and the inability to know what happens in those black boxes, has given rise to the field of Explainable Artificial Intelligence, XAI. Ribeiro et al. (2016) showed that when an AI model was asked to explain why it had identified a dog as a wolf, it highlighted the pixels in the picture that showed a snowy background. Why did that happen? Well, because no matter how good the AI model is it does not *know* what a dog is, nor a wolf. It only identifies patterns that exist in images that humans say contain a dog or a wolf. The classification is based more on human selection (of images, not of dog breeds) than on the abilities of the algorithm.

This is one of the fundamental problems that feeds into O'Neil's critique of algorithms—they are powerful statistical engines that some decisions-makers rely heavily on, without understanding that the machines do not actually *know* anything and cannot tell when a decision is based on bad or insufficient data. Tversky and Kahneman (1974) also show that humans are bad at making decisions, and bad at justifying them, but it is easier second guessing a human than a statistical engine.

London (2019) and Zerilli et al. (2019) both base their skepticism towards the kinds of explanations required of algorithms, in contrast to the expectations on human decisions. London looks at decisions and predictions in the medical sciences, and those decisions compare to what is required of an AI in the same field. For many medical conditions, and regarding many medicines, the medical sciences simply do not understand the causal effects. That taking a certain drug most often leads to a certain effect can be shown, but not by which causal process that happens. Still, people trust doctors and the medical sciences. Why, then, should the medical sciences require more advanced explanations from an AI than from a human? In fact, Wischmeyer (2020) goes so far as to say that in terms of legislation, there is no need to require thorough explanations from AI systems, as the law is already accustomed to the unspecific justifications of human decision-making.

While these are reasonable arguments, there are at least two objections to the above position. The first argues that an AI *should* be subject to harsher requirements than humans, because of the higher processing power and capacity for comparison in an AI compared to humans (Siau and Wang, 2018). Therefore, an AI should be subject to harsher requirements when explaining a decision. The other objection is evident in the papers included in this thesis. For the companies

studied in Papers II, III, and IV, there is not a single company that has shown that they use a black box AI. The evidence from the sample is that the technology is not used to the extent that many fear or believe. At least not in companies that are not big tech, or tech start-ups.

An adjacent problem is the overly burdensome requirements set on customers' ability to understand advanced algorithms (and some not so advanced ones). This was shown through the use of consumer requests for explanations in Papers III and IV. The hurdles identified in Paper IV present at least two such instances. The first, *expertise requirements*, shows that if a customer receives a vague or unsatisfactory response to a request for information (which seems highly likely given the sample), it requires some technical expertise to be able to question the response. Should the average customer have to understand how machine learning works in order to receive a decent response? Or must they even know the specifics about the algorithms a company uses? That does not seem to be the intention of the legislator, but it might still be a consequence of the legislation. A related hurdle is *targeted language*, which indicates that the customer needs to use language specifically tied to the services the company uses or provides in order for the customer service representative to recognize that the request is relevant. Quoting the rights set out in the legislation does not seem to be enough, something Sørum and Presthus (2020) also recognized.

Finally, an important note to finish this section on concerns Winner's 1980 assertion that "artifacts have politics". What Winner meant by this is that there is no such thing as a neutral technology. An industrial loom, of Luddite fame, is not neutral; it shifts the power dynamic of a business away from manual labor to industrial labor—from the worker to the owner. In this sense, algorithms are not neutral. They can shift power towards the few, or towards the many. They can reinforce stereotypes or reduce their impact. The problem is that it is hard to make those choices, because it is not known how well the data represents the world (Andrus et al., 2021), and because the technologies are opaque. As an example, Dignum (2022) suggests adopting a whole new approach to AI development to remedy this, namely Relational AI.

Is it any wonder, when the technologies are opaque, unexplainable and the legal requirements are unclear yet burdensome, that companies simply avoid using AI technologies in their service? This neatly leads to another hurdle to transparency: *Avoidance*.

Explaining things is generally difficult

The third and final reason for why transparency is difficult to achieve is that it is more difficult to provide adequate explanations than one might initially assume.

In the literature there are several examples of researchers trying to design better explanations, either for computer systems in general or for automated decision-making in particular. These efforts have in common that the authors really want to improve transparency, they want to create pedagogical examples,

and create solutions that can later inspire companies and other actors to improve their transparency.

The problem is that the examples fall into roughly two categories. Either the explanations designed do not appear to get users to an appropriate level of understanding (Sadeghi et al., 2021; Rader et al., 2018), or the explanations are more elaborate than a company is likely to apply to their services (Binns et al., 2018; Ehsan et al., 2021; Cheng et al., 2019; Hamon et al., 2021). This research shows that there is a difficult trilemma to negotiate—an explanation needs to be accurate in how it describes the system. It must be sufficiently detailed to provide the user with enough insight to make an informed choice, sufficiently simple for most users to understand it, and sufficiently easy for a company to apply for all the different points in a system that need to be explained. On a promising note, Bove et al. (2022) do show an explanation design that creates a positive relationship between transparency and satisfaction, and their explanations seem to negotiate the trilemma in the context of insurance pricing.

A further problem is what Simon (1971) discussed: "a wealth of information creates a poverty of attention". This sentiment was also represented in Paper II; there is too much information available for people to be able to take it in, and if reading through the entirety of a terms of service-agreement for each service, a user signs simply takes on too much. If explanations are to work, then a user needs to be able to pay attention to that explanation.

Dexe et al. (2020) and Papers III and IV also showed that the responses received were lackluster. However, Keil (2006) and Wilson and Keil (1998) would likely defend the appropriateness of those explanations to a degree. If there is a general lack of attention, then *shallow* explanations might not be a bad idea. Responses can be analyzed through the *explanatory domain* of insurance, in that the customer who has made the request knows that the answer received is in the domain of insurance. The idea of *causal power* can be applied to say that information in the explanatory domain of insurance likely behaves in a certain way, for instance by increasing or lowering risk, thereby increasing or lowering price. The explanations might not stand up to legal scrutiny, but they may be much more appropriate for most users than a researcher dissecting the answers, as initially assumed.

Papers III and IV showed that the hurdle *legal interpretations* further complicates how automated decision-making can be explained. This, in turn, sets the stage for two additional hurdles: the need to be able to specify exactly what information is being requested (*specificity*), even going so far as having to explain the details of the law to the company, and, if the company has a different interpretation of the law, a further reduction of the consumers' *ability to question* a decision made by the automated system.

As mentioned in the beginning of this section, there is, of course, the scenario where a customer receives a response, and it seems to fit with what they asked for, but answer is just not very good. There are no traces of any of these types of explanations, which Keil (2006) lists. All are simply *bad explanations*.

5.3 Usability and design

Facing the problems described in section 5.2, and still wanting to realize the benefits described in section 5.1, the question of what to do instead comes to the forefront. How does one realize the benefits of transparency while negotiating both the possible drawbacks, as well as the balance between what should be disclosed and what should not? As alluded to in both sections 1.2 and 2.5, design and usability might be one way to realize such benefits.

It was shown, through the ISO definitions of usability and satisfaction as well as Norman (1988) and Bardzell and Bardzell (2015), that there is a need to align designer intentions, the user interfaces and the mental models of the user in order to make sure that users are satisfied with a system. Making explicit what the designer wants the system to do, how the interface accomplished this and how it aligns with user perceptions is transparency and explanations in action.

Being more explicit with what a system does, and being transparent with the system's functionality and consequences, are ways to bridge the gaps. Keil (2006) noted a difference between mental models and explanations. Mental models are "readouts of relations from a mental array and are often understood in spatial terms". Explanations, Keil notes, are not simply descriptive blueprints or plans but are also interpretations of those plans. Can satisfaction be achieved simply by the fact that a mental model aligns with how the system architecture looks? It can probably be deduced that satisfaction not only requires an alignment of the mental model of a system, but also the *interpretation* of that system. Ergo, satisfaction requires that the system is explained, not simply mapped out. It requires that designers are able to give users the right expectations of the product, as was pointed out in Paper II.

This has been a long-winded way of showing that for a design to fulfill the requirements of usability, primarily the requirements of satisfaction with a product, that design also needs an element of transparency. Kizilcec (2016) even showed that not being transparent can seriously reduce satisfaction with a service if expectations are violated (a negative decision when users expected a positive one). However, if instead you explain the reasons for why a negative decisions was taken—if you take steps to align with the users' mental model—that loss of satisfaction can be negated through transparency. Bove et al. (2022) also show that transparency and explanations can increase satisfaction with an insurance purchase.

Nevertheless, does transparency actually lead to understanding? Section 5.2 seems to provide ample evidence that transparency has its flaws but transparency used strategically and with intention means the answer might very well be 'yes'.

Section 2.1 discussed transparency as a *pro-ethical* condition, a thing that can enable or impair other ethical values (Turilli and Floridi, 2009), which was shown in Figures 2.1 and 2.2. Certain ethical values are dependent on transparency in order to be realized, such as accountability and informed consent. Other values require regulating transparency in order to be realized, such as privacy and

autonomy. Since transparency is not an intrinsic good, it does not create benefits just by existing, and Turilli and Floridi (2009) also show that for transparency to realize ethical values, the designers need to apply transparency with 'direction' or 'intentionality'. Transparency is a tool that can be used to accomplish a goal.

Several researchers also point this out, directly or indirectly. Even though Lessig (2009) argues for why transparency may be ill advised, the point of the text is not to say that transparency is not important, but that governments, and other actors, need to be aware of the negative consequences of their choices: "Reformers rarely feel responsible for the bad that their fantastic new reform effects. Their focus is always on the good. The bad is someone else's problem" (Lessig, 2009). Thinking about the negative consequences of transparency is akin to needing to think about what a specific type of transparency accomplishes, and if the 'right' transparency is applied in a given situation.

Hutchinson et al. (2021) argue that not thinking properly about how data sets are constructed creates problems with accountability. This point is similar to Andrus et al. (2021) who looks at the problems with discrimination that might come from not thinking about the consequences of using demographic data. Both are aligned with how Turilli and Floridi (2009) argue that even the creation of data is something with which actors need to be transparent, and that ethical values, like accountability or equality, also depend on the way actors explain how information is created from data.

This section began by discussing the link between usability and explanations, as well as transparency. Keil (2006) argued that there is a difference between a mental model and explanations in that explanations also include an interpretation of the model. One of the reasons for this, Keil argues, is that explanations are transactional. Explanations have a sender, a company explaining an automated decision-making system, and a receiver, a user interpreting what the automated decision-making system actually does.

Knowing that explanations are transactional, and that transparency requires purpose, it can be deduced that an important contribution to improving transparency should come from interaction design, and human computer interaction in general. Paper IV argued that the very first hurdle is the strategy by which companies have chosen to be transparent, namely their *interaction strategy*.

5.4 Limitations

Several limitations of the papers included in this thesis are mentioned in the papers themselves. Thus, this section covers a few general limitations concerning the overarching thesis work.

It is evident that a limitation common to all papers included is the generalizability of the results. Paper III has the best coverage of the areas studied, covering from 40 to 95% of the respective insurance markets. However, it still only covers a single insurance product and only a single type of business. Paper

II has a decent sample for an interview study with experts, but it consists of only a small fraction of insurance professionals in Sweden and does not represent a majority of insurance companies. Paper I could have included more samples from the at least some of the included countries and could have combined the results with interviews of government coordinators or investigators in the field. Paper IV is, in some ways, a study about the problems with data collection, but much more effort could have been made in reaching out to other companies.

That being said, after four studies and four years during which I and my co-authors studied these topics, the samples included give a good idea of what the implementation of transparency looks like across the industries studied, both in terms of how the data is motivated and in terms of the general impression, after having talked to companies and organizations about the studies.

Another point that is relevant to make in the context of generalisations is that the studies have not come close to studying *all* the information that companies have available about products, services, ethical principles, policies, comments on legislation, annual reporting and corporate culture, or any other piece of information that could reasonably inform customers about what a company does and does not do. The fact is that most customer facing companies have a large array of information available in different ways. This is not always about automated decision-making, at least not more than what the law requires, but it is still information that can produce trust, positive attitudes and increase satisfaction.

None of the papers include user studies, a staple of HCI-research. Neither have the authors designed any alternative explanations (although drafts have been made and discarded). Had they done so, results could have come closer to verifying the needs and wants of users in terms of transparency, or the understandability of explanations.

It is, however, unlikely that there are appropriate standardized solutions for discovering how to explain algorithms. The literature reviews and results seem to indicate that transparency has more to do with leaning into complexity. Companies are likely forced to think about how to be transparent with a specific product, for specific purposes, towards a specific audience. What has been provided, instead, are ways to think about transparency in order to make those adaptations possible.

Finally, the topic at hand is by its nature multi-disciplinary. However, even if I, and my co-authors, have experiences in other scientific disciplines to add to the understanding of transparency in an HCI-context, the thesis also touches on a number of highly developed scientific fields in which neither author are specialists. This thesis has covered discussions on law, sociology of law, philosophy, psychology, computer science, and political science, and those are just the fields where the discussions were written down in this thesis. The references cover a broader span still. I hope that I have done those fields justice, and that they serve to inform the field of human-computer interaction, but I am by no means an expert in either of those fields, and as such it is not for me to say whether I have succeeded.

5.5 Future Research

This section contains suggestions for future research that could either strengthen the theories included, or challenge them.

First of all, as has been noted, there is a need for more research into how transparency can be used to realize different values. Some attempts have been presented already, with experimental designs aimed at perceptions of justice (Binns et al., 2018) or user satisfaction (Bove et al., 2022). As this thesis has established through Turilli and Floridi (2009), transparency is a *pro-ethical* condition that can be used to realize different ethical values. As Binns et al. and Bove et al. have done, there is a need to investigate this further in experimental settings, and in contextual settings. How, exactly, does transparency affect accountability in, say, financial reporting? To what extent should information transparency be regulated in order to maintain an individual's sense of appropriate privacy? The experimental settings, mixed with real world examples or data, would ideally lay the groundwork for more testing in corporate or governmental services.

In addition to the experimental settings, more research is needed into the practical adoptions of transparency—finding out what information consumers actually have access to and how that informs their opinion and acceptance of various services. As shown through the various versions of Delade meningar there is a lot of skepticism towards data collection and of various data processing practices. Sørum and Presthus (2020), Dexe et al. (2020) and Paper III and IV all serve to investigate these practices, but more effort is needed in mapping out what the customer experience of transparency actually is.

Since many algorithms used by various companies, governments and non-governmental organizations are not developed in-house, they are often purchased from some external party, more research is needed into how organisation can formulate requirements for transparent and explainable algorithms. This can be done in different contexts as well. One prominent and important arena would be public procurement. Governments, public authorities and local government are different beasts than many companies. They serve a different audience, with a different mandate. As such, they also have particular demands on algorithms that they need to be able to express, and that companies they procure systems from need to be able to understand and set into practice. Identifying how the values a public administration wants to realize can be realized through transparency, and how they can formulate requirements to make sure they are realized is an important field to study further. Requiring explanations from AI systems is the present day equivalent to how difficult it was to procure usability earlier (Artman and Zällh, 2005; Artman et al., 2010).

Another context in which such procurements can be investigated is when companies purchase products (algorithms) from each other. Since each company is responsible for the technology it uses, and in part how others use their technology, and that being transparent to customers requires that companies understand the products they in turn are using, requires that the different actors in a business-

to-business purchase of algorithms understand each other. They need to be able to formulate contexts in which the algorithms are to be applied, and to formulate the risks inherent in different technological solutions. Following such B2B relationships would be highly interesting research for several fields.

Furthermore, research is required in using explanations regarding algorithmic decisions that are not based primarily on legally complient texts, but rather on the more "naturalistic" explanations shown in Keil (2006) and Wilson and Keil (1998).

For the field of HCI in general, it would be interesting to see the development of a design theory for transparency. One that can be taught to practitioners, or future practitioners, to make sure that transparency and explanations are not an afterthought, but are instead an essential part of creating trustworthy, user friendly and usable designs.

Finally, more research is needed into the negative and false transparency that has been mentioned throughout this thesis. Being able to show how and where transparency fails is as important for being able to make use of it in beneficial ways as is showing incremental positive effects.

Chapter 6

Conclusions

Three questions were asked in the beginning of this thesis:

RQ1 What benefits does information transparency generate?

RQ2 Why is it so hard to achieve transparency in automated decision-making?

RQ3 In what ways does transparency relate to usability?

They have, throughout this thesis, been investigated on a theoretical and practical level. Available research has been consulted, experts interviewed and data collected through consumer rights.

Transparency seems to be more enticing as an abstract goal than a concrete measure. It is a well-established tool in order to fight corruption and other financial wrong doing (Bauhr and Grimes, 2014; Hood and Heald, 2006). It is desirable among the public (Delade meningar, 2019) and several researchers have shown how it can increase trust services (Kim and Lee, 2012; Kang and Hustvedt, 2014; Kim and Kim, 2017; Bhaduri and Ha-Brookshire, 2011; Cambier and Poncin, 2020). It has also been shown in secion 1.1 that there is a desire to have more transparency in general, and that accountability and informed consent are almost impossible without transparency (Turilli and Floridi, 2009).

However, despite these advantages, there is little evidence that it is easy to achieve these benefits through transparency. If it were, then the calls for increased transparency might be less important. The question should perhaps be *how* to realize the benefits of transparency, rather than if there are any.

Even getting to the how requires dealing with a few hurdles.

First of all, transparency is not always the right choice. In certain situations transparency can impair ethical values society (or specific organisations) wants to uphold (Turilli and Floridi, 2009). Business confidentiality and competition are two reasons for why companies would not want to be transparent with all of their information. It is also the case that companies, and other organizations for that matter, must face the risk that it is hard to control how the public, or other

organizations, interpret the information with which the company is transparent. Allan and Berild Lundblad (2021b) note that this is a reason for why companies choose not to be transparent: they might not know how to control the story.

Second, technological advancements make transparency harder to accomplish. Black box models are unexplainable, and even with advanced models that help such systems produce explanations, those explanations hard to verify, and humans may place unwarranted trust in them (Lakkaraju and Bastani, 2020). London (2019) and Zerilli et al. (2019) both argue that perhaps policy makers should consider not placing such harsh requirements on to what extent an AI can explain itself, because, as they argue, humans are not that good at explaining their own decisions (Tversky and Kahneman, 1974). Siau and Wang (2018) disagrees, arguing that the fact that an AI exceeds humans in certain cognitive aspects means that society should place even harsher requirements on AI explanations than on human ones.

Black box systems aside, algorithms may still be harder to explain for most people than both legislators and computer scientists expect. This was shown in Paper III and IV, as in Sørum and Presthus (2020).

Why? Well, that is the third point. It might be harder to explain things than most assume. Efforts to design better explanations for automated decision-making are either not good enough to get users to understand the systems (Sadeghi et al., 2021; Rader et al., 2018) or are too advanced and resource intensive to be easily adopted by companies (Binns et al., 2018; Ehsan et al., 2021; Cheng et al., 2019; Hamon et al., 2021), even in the few cases where the solutions modeled give clear effects.

The only exception was shown by Bove et al. (2022) where results indicated a clear relationship between transparency and satisfaction, and the explanations seem appropriate for the given context. However, even then, designers need to make sure the users can pay adequate attention (Simon, 1971).

Keil (2006) and Wilson and Keil (1998) argue that most every day explanations are much shallower and simpler than what the explanation and transparency literature often assumes. By means of cognitive shortcuts such as explanatory domains and causal patterns, humans are able to understand information in context, something which out of context would be incomprehensible. However, such explanations are not what researchers look for, and the interpretive and critical stance most researchers take might fail to appreciate the effectiveness and appropriateness of the simple contextual explanation.

Having sorted out the many problems with realizing the benefits of transparency, the thesis then sought out to understand what designers and practitioners can do to actually use transparency for good.

Usability is a central tenant in Human-Computer Interaction. By consulting the international standards defining usability (ISO, 2018) it was established that satisfaction with a product is a requirement for achieving usability. Satisfaction, in turn, requires that the mental models a user has of a product align with how the product actually works—that is, it depends on the expectations users have. If

those do not align (and do so in a negative way) then the use of the product will cause discomfort and negative attitudes. Thus, in order to make sure that users are satisfied with a product, which in turn affects usability, user expectations need to be managed.

Transparency is an excellent way to manage expectations, especially if that transparency is also designed with the intentionality to realize certain ethical values and specific goals (Turilli and Floridi, 2009). Moreover, if the information with which companies are transparent include a measure of explanation that increase the users understanding of what is being disclosed, then explanations are not simply descriptions of a thing (Keil, 2006), they also include the recipients interpretations of the information. User testing, designing prototypes and evaluating transparency continuously become important tools to realize the benefits of transparency.

Transparency is not neutral (Turilli and Floridi, 2009; Winner, 1980). Using it requires that companies and other organizations think about the consequences of the transparency, and what goals it aims to achieve. It might be equally important to be open and explicit with the limitations of your transparency, and to be comfortable with admitting that in certain aspects transparency can be counterproductive, as it is to provide appropriate explanatory value to the information. As Larsson and Heintz (2020) point out, transparency has to be understood in its applied context.

The papers included in this thesis paint a fairly dark picture of how transparent companies tend to be. They do not agree on the benefits or the practice, and even if they happen to respond to requests for information, those responses leave plenty to be desired.

Nevertheless, it is also the case that the thesis has only explored a small amount of information in a large and complex eco-system. There are plenty of explanations and descriptions of products, practices, values, policies and social responsibilities that are available at websites, in contracts, in advertisement and through customer service representatives.

Most companies are transparent. Most customers have much information to access. There is more to do in terms of making sure that customers have access to the right information, at the right time, on the right level of complexity. This thesis is an attempt to increase the awareness of how to think about how to make those choices and improve transparency even more. Just because something is transparent that does not mean that it is comprehensible. You can be both transparent and incomprehensible.

The need to strategically think about transparency is important. If companies are able to figure out what values they want to adhere to and promote, if companies are able the set the correct expectations in a user, and if companies are able to make sure that explanations accurately and effectively portray what they want to accomplish, then transparency can be an effective tool to achieve those benefits.

6.1 Contributions

Paper	Type of contribution	Description of contributions
Paper I	Case study	A study of whether ethical AI leads to competitive advantages, using government AI strategies
	Framework	Provides a framework for analyzing how government AI strategies promote ethical practices
Paper II	Rich insight	A description of a contextualized understanding of transparency in the context of insurance
	Case study	Investigating how insurance experts view the benefits and drawbacks of transparency in insurance
	Proposition	Identified commonalities and limitations of transparency in a specific context
Paper III	Case study	How insurance companies respond to right to access requests from consumers
	Research Method	Development of method using consumer rights to get access to data about information practices
	Framework	A description of limitations and justifications of GDPR implementations in insurance
Paper IV	Case study	How companies, broadly, respond to right to access requests from consumers
	Proposition	There are clear limitations to the practice of transparency in algorithmic decision-making
	Generative Mechanism	Description of nine hurdles to transparency

Table 6.1: Theoretical contributions of the papers based on Presthus and Munkvold (2016)

This thesis and its constituent papers make several different scientific contributions. Presenting information in a long form text does, however, sometimes make it difficult to say exactly which different parts contribute. Taking the advice of Presthus and Munkvold (2016), this section provides an attempt to systematize the theoretical contributions made. Table 6.1 details the different papers included in this thesis, and Table 6.2 lists the contributions of the thesis as a whole. Descriptions of the types of contributions can be found in Presthus and Munkvold (2016).

Considering the model for transparency and explanations presented in Figure 2.3, it is also worth while to map the constituent papers of the thesis according to this model. These are presented in Figure 6.1. Paper I aimed at investigating a link between ethical AI and a competitive advantage, but no such concrete connection was found. In Paper II, instead, the terminology shifted from ethics to

Type of contribution	Description of contributions
Rich insight	A description of transparency based on established research and our own case studies
Framework	Thorough description of how transparency, explanations and usability function within the context of algorithmic decisions
Model	Details how transparency, explanations and usability affect each other when designing and explaining systems
Mid-range theory	Brings theories of transparency and explanations to the field of HCI, and suggests how HCI could use these theories

Table 6.2: Theoretical contributions of the overall thesis based on Presthus and Munkvold (2016)

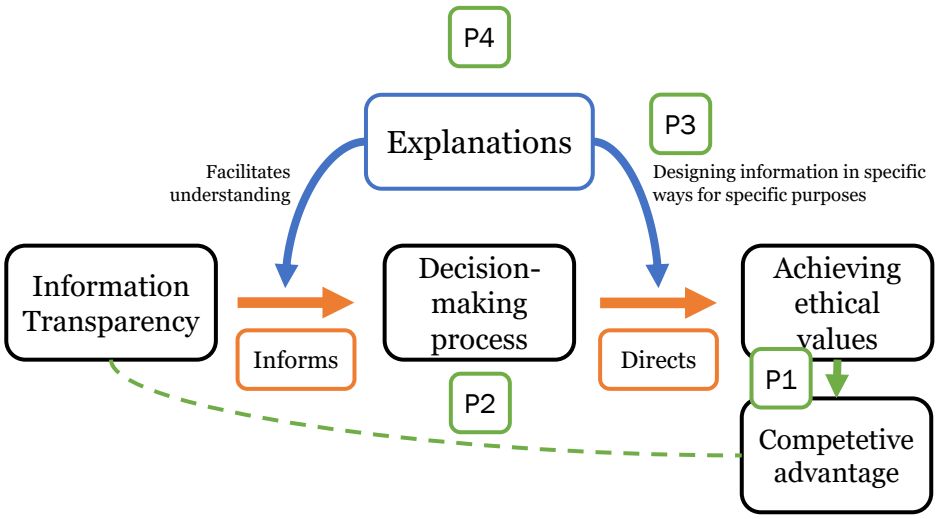

Figure 6.1: Contributions of each paper of the thesis mapped to the model for transparency and explanations.

transparency, but the point of interest was still the realization of a competitive advantage. Compared to Paper I, Paper II narrowed the scope significantly to a single industry (insurance), a single country (Sweden), and another group of actors (insurance professionals). While the insurance professionals all saw advantages with transparency, there was a lack of agreement as to what it meant, and what processes should be made transparent. Based on this lack of ideas concerning practical applications, Paper III instead sought to see if the problem lay with how the information can be explained. It focused on the same industry, with the

same terminology but looked instead at practical applications of transparency requirements in the GDPR across different countries. In Paper IV, the scope was changed somewhat (looking at several industries instead of one), but the question remained. However, due to the lack of replies and lack of transparency, the paper instead tried to explain the hurdles that are in the way of transparency, with a heavy focus on the role that explanations play.

References

Abdul A, Vermeulen J, Wang D, Lim BY, Kankanhalli M (2018) Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda. In: Proceedings of the 2018 CHI conference on human factors in computing systems, pp 1–18, DOI 10.1145/3173574.3174156

Abiteboul S, Stoyanovich J (2019) Transparency, fairness, data protection, neutrality: Data management challenges in the face of new regulation. Journal of Data and Information Quality 11(3), DOI 10.1145/3310231

Adebayo J, Gilmer J, Muelly M, Goodfellow I, Hardt M, Kim B (2018) Sanity checks for saliency maps. In: Advances in Neural Information Processing Systems, vol 31, p 9525–9536, URL https://proceedings.neurips.cc/paper/2018/file/294a8ed24b1ad22ec2e7efea049b8737-Paper.pdf

Agre PE, Rotenberg M (eds) (1997) Technology and Privacy: The New Landscape. MIT Press, Cambridge, MA, USA

Ahmad K (2021) Human-centric requirements engineering for artificial intelligence software systems. In: 2021 IEEE 29th International Requirements Engineering Conference (RE), pp 468–473, DOI 10.1109/RE51729.2021.00070

Alizadeh F, Jakobi T, Boden A, Stevens G, Boldt J (2020) GDPR reality check—claiming and investigating personally identifiable data from companies. In: 2020 IEEE European Symposium on Security and Privacy Workshops (EuroS PW), pp 120–129, DOI 10.1109/EuroSPW51379.2020.00025

Allan R, Berild Lundblad N (2021a) Regulate tech #14: Classics - Lessig and Code. URL https://regulatetech.podbean.com/e/regulate-tech-14-classics-lessig-and-code/

Allan R, Berild Lundblad N (2021b) Regulate tech #29: Transparency - how, who, and for what? URL https://regulatetech.podbean.com/e/regulate-tech-29-transparency-how-who-and-for-what/

Alvarado O, Waern A (2018) Towards algorithmic experience: Initial efforts for social media contexts. In: Proceedings of the 2018 CHI Conference on Human

Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, CHI '18, p 1–12, DOI 10.1145/3173574.3173860

Andrews R, Diederich J, Tickle AB (1995) Survey and critique of techniques for extracting rules from trained artificial neural networks. Knowledge-based systems 8(6):373–389, DOI 10.1016/0950-7051(96)81920-4

Andrus M, Spitzer E, Brown J, Xiang A (2021) What we can't measure, we can't understand: Challenges to demographic data procurement in the pursuit of fairness. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 249–260, DOI 10.1145/3442188.3445888

Appelgren E (2017) The reasons behind tracing audience behavior: A matter of paternalism and transparency. International Journal of Communication 11:20

Appelgren E, Leckner S (2016) Att dela eller inte dela: internetanvändarnas inställning till insamling av personliga data. SOM-institutet

Artman H, Zällh S (2005) Finding a way to usability: Procurement of a taxi dispatch system. Cognition, Technology and Work 7(3):141–155, DOI 10.1007/s10111-005-0182-6

Artman H, Dovhammar U, Holmlid S, Lantz A, Lindquist S, Markensten E, Swartling A (2010) Att beställa något användbart är inte uppenbart: En motiverande bok om att beställa användbarhet

Asadabadi MR, Chang E, Zwikael O, Saberi M, Sharpe K (2020) Hidden fuzzy information: Requirement specification and measurement of project provider performance using the best worst method. Fuzzy Sets and Systems 383:127–145, DOI 10.1016/j.fss.2019.06.017

Ashby S, Hanna J, Matos S, Nash C, Faria A (2019) Fourth-wave HCI meets the 21st century manifesto. In: Proceedings of the Halfway to the Future Symposium 2019, Association for Computing Machinery, New York, NY, USA, HTTF 2019, DOI 10.1145/3363384.3363467

Bahşi H, Franke U, Langfeldt Friberg E (2019) The cyber-insurance market in Norway. Information and Computer Security 28(1):54–670, DOI 10.1108/ICS-01-2019-0012

Baldwin R, Scott C, Hood C (1998) A Reader on Regulation. Oxford Readings in Socio-Legal Studies, Oxford University Press, Oxford

Ball C (2009) What is transparency? Public Integrity 11(4):293–308, DOI 10.2753/PIN1099-9922110400

Bannon LJ (1995) From human factors to human actors: The role of psychology and human-computer interaction studies in system design. In: Readings in human–computer interaction, Elsevier, pp 205–214, DOI 10.1016/B978-0-08-051574-8.50024-8

Barakat N, Bradley AP (2010) Rule extraction from support vector machines: a review. Neurocomputing 74(1-3):178–190

Barclay I, Abramson W (2021) Identifying Roles, Requirements and Responsibilities in Trustworthy AI Systems, Association for Computing Machinery, New York, NY, USA, p 264–271. URL 10.1145/3460418.3479344

Bardzell J, Bardzell S (2015) Humanistic HCI. Synthesis Lectures on Human-Centered Informatics Series, Morgan & Claypool Publishers, DOI 10.2200/S00664ED1V01Y201508HCI031

Bauhr M, Grimes M (2014) Indignation or resignation: The implications of transparency for societal accountability. Governance 27(2):291–320, DOI https://doi.org/10.1111/gove.12033

Benjamin R (2019) Race After Technology: Abolitionist Tools for the New Jim Code. John Wiley & Sons

Berild Lundblad N (2018) What are we talking about when we talk about algorithmic transparency? URL https://anteckningarna.org/2017/10/31/what-are-we-talking-about-when-we-talk-about-algorithmic-transparency/, accessed on February 25, 2022.

Bevan N (2010) Extending the concept of satisfaction in iso standards. In: Proceedings of the KEER 2010 International Conference on Kansei Engineering and Emotion Research

Bevan N, Carter J, Harker S (2015) ISO 9241-11 revised: What have we learnt about usability since 1998? In: International conference on human-computer interaction, Springer, pp 143–151, DOI 10.1007/978-3-319-20901-2_13

Bhaduri G, Ha-Brookshire JE (2011) Do transparent business practices pay? Exploration of transparency and consumer purchase intention. Clothing and Textiles Research Journal 29(2):135–149, DOI doi.org/10.1177/0887302X11407910

Bibal A, Lognoul M, de Streel A, Frénay B (2021) Legal requirements on explainability in machine learning. Artificial Intelligence and Law 29(2), DOI 10.1007/s10506-020-09270-4

Binns R, Van Kleek M, Veale M, Lyngs U, Zhao J, Shadbolt N (2018) 'It's reducing a human being to a percentage': Perceptions of justice in algorithmic decisions. In: Proceedings of the 2018 CHI Conference on human factors in computing systems, ACM, CHI '18, pp 1–14, DOI 10.1145/3173574.3173951

Bødker S (2006) When second wave HCI meets third wave challenges. In: Proceedings of the 4th Nordic conference on Human-computer interaction: changing roles, pp 1–8, DOI 10.1145/1182475.1182476

Bødker S (2015) Third-wave HCI, 10 years later—participation and sharing. interactions 22(5):24–31, DOI 10.1145/2804405

Bottis M, Panagopoulou-Koutnatzi F, Michailaki A, Nikita M (2019) The right to access information under the GDPR. International Journal of Technology Policy and Law 3(2):131–142, DOI 10.1504/IJTPL.2019.104950

Bove C, Aigrain J, Lesot MJ, Tijus C, Detyniecki M (2022) Contextualization and exploration of local feature importance explanations to improve understanding and satisfaction of non-expert users. In: 27th International Conference on Intelligent User Interfaces, Association for Computing Machinery, New York, NY, USA, IUI '22, p 807–819, DOI 10.1145/3490099.3511139

Bradford A (2020) The Brussels Effect: How the European Union Rules the World. Oxford University Press, DOI 10.1093/oso/9780190088583.001.0001

Brandeis LD (1913) What publicity can do. Harpers's Weekly 58(2974):10

Brkan M, Bonnet G (2020) Legal and technical feasibility of the GDPR's quest for explanation of algorithmic decisions: of black boxes, white boxes and fata morganas. European Journal of Risk Regulation 11(1):18–50, DOI 10.1017/err.2020.10

Butcher J, Beridze I (2019) What is the state of artificial intelligence governance globally? The RUSI Journal 164(5-6):88–96, DOI 10.1080/03071847.2019.1694260

Cambier F, Poncin I (2020) Inferring brand integrity from marketing communications: The effects of brand transparency signals in a consumer empowerment context. Journal of Business Research 109:260 – 270, DOI 10.1016/j.jbusres.2019.11.060

Chazette L, Brunotte W, Speith T (2021) Exploring explainability: A definition, a model, and a knowledge catalogue. In: 2021 IEEE 29th International Requirements Engineering Conference (RE), pp 197–208, DOI 10.1109/RE51729.2021.00025

Cheng HF, Wang R, Zhang Z, O'Connell F, Gray T, Harper FM, Zhu H (2019) Explaining Decision-Making Algorithms through UI: Strategies to Help Non-Expert Stakeholders, Association for Computing Machinery, New York, NY, USA, p 1–12. DOI 10.1145/3290605.3300789

Comber R, Lampinen A, Haapoja J (2019) Towards post-interaction computing: Addressing immediacy, (un)intentionality, instability and interaction effects. In: Proceedings of the Halfway to the Future Symposium 2019, pp 1–8, DOI 10.1145/3363384.3363477

Cormen T, Leiserson C, Rivest R, Stein C (2009) Introduction to Algorithms. Computer science, McGraw-Hill

Cysneiros LM, do Prado Leite JCS (2020) Non-functional requirements orienting the development of socially responsible software. In: Nurcan S, Reinhartz-Berger I, Soffer P, Zdravkovic J (eds) Enterprise, Business-Process and Information Systems Modeling, Springer International Publishing, Cham, pp 335–342, DOI 10.1007/978-3-030-49418-6_23

Dafoe A (2018) AI governance: a research agenda. Future of Humanity Insitute URL `https://www.fhi.ox.ac.uk/wp-content/uploads/GovAIAgenda.pdf`, accessed on March 13, 2022.

Damro C (2012) Market power Europe. Journal of European Public Policy 19(5):682–699, DOI 10.1080/13501763.2011.646779

de Laat PB (2018) Algorithmic decision-making based on machine learning from big data: Can transparency restore accountability? Philosophy & Technology 31(4):525–541, DOI 10.1007/s13347-017-0293-z

de Vries K (2010) Identity, profiling algorithms and a world of ambient intelligence. Ethics and Information Technology 12(1):71–85, DOI 10.1007/s10676-009-9215-9

Delade meningar (2019) Delade meningar 2019 [Shared opinions 2019]. Tech. rep., Insight intelligence, URL `https://www.insightintelligence.se/delade-meningar/delade-meningar-2019`

Delade meningar (2020) Delade meningar 2020 [Shared opinions 2020]. Tech. rep., Insight intelligence, URL `https://www.insightintelligence.se/delade-meningar/delade-meningar-2020`

Delade meningar (2021) Delade meningar 2021 [Shared opinions 2021]. Tech. rep., Insight intelligence, URL `https://www.insightintelligence.se/delade-meningar/delade-meningar-2021/`

Delade meningar (2022) Delade meningar 2022 [Shared opinions 2022]. Tech. rep., Insight intelligence, URL `https://www.insightintelligence.se/delade-meningar/delade-meningar-2022/`

Dexe J, Franke U (2020) Nordic lights? National AI policies for doing well by doing good. Journal of Cyber Policy 5(3):332–349, DOI 10.1080/23738871.2020.1856160

Dexe J, Ledendal J, Franke U (2020) An empirical investigation of the right to explanation under GDPR in insurance. In: Trust, Privacy and Security in Digital Business. The 17th International Conference on Trust, Privacy and Security in Digital Business – TrustBus 2020, Springer, DOI 10.1007/978-3-030-58986-8_9

Dexe J, Franke U, Rad A (2021) Transparency and insurance professionals: a study of swedish insurance practice attitudes and future development. The Geneva Papers on Risk and Insurance–Issues and Practice 46(4):547–572, DOI 10.1057/s41288-021-00207-9

Dexe J, Franke U, Söderlund K, van Berkel N, Jensen RH, Lepinkäinen N, Vaiste J (2022) Explaining automated decision-making: a multinational study of the GDPR right to meaningful information. The Geneva Papers on Risk and Insurance–Issues and Practice pp 1–29, DOI 10.1057/s41288-022-00271-9

Dignum V (2022) Relational artificial intelligence. DOI 10.48550/arXiv.2202.07446, 2202.07446

Dobrev D (2003) A definition of artificial intelligence. Mathematica Balkanica 19:67–74

Du M, Liu N, Hu X (2019) Techniques for interpretable machine learning. Communications of the ACM 63(1):68–77, DOI 10.1145/3359786

Duarte EF, Baranauskas MCC (2016) Revisiting the three HCI waves: A preliminary discussion on philosophy of science and research paradigms. In: Proceedings of the 15th Brazilian Symposium on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, IHC '16, DOI 10.1145/3033701.3033740

Ehsan U, Liao QV, Muller M, Riedl MO, Weisz JD (2021) Expanding explainability: Towards social transparency in AI systems. Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems DOI 10.1145/3411764.3445188

Elmer G (2003) Profiling Machines: Mapping the Personal Information Economy. MIT Press

Esaiasson P, Gilljam M, Oscarsson H, Wängnerud L (2007) Metodpraktikan. Konsten att studera samhälle, individ och marknad 3(1):12–19

Eubanks V (2018) Automating Inequality: How High-Tech Tools Profile, Police, and Punish the Poor. St. Martin's Publishing Group

Fan M, Yu L, Chen S, Zhou H, Luo X, Li S, Liu Y, Liu J, Liu T (2020) An empirical evaluation of GDPR compliance violations in android mHealth apps. In: 2020 IEEE 31st International Symposium on Software Reliability Engineering (ISSRE), pp 253–264, DOI 10.1109/ISSRE5003.2020.00032

Filimowicz M, Tzankova V (2018) Introduction— new directions in Third Wave HCI. In: New Directions in Third Wave Human-Computer Interaction: Volume 1-Technologies, Springer, pp 1–10, DOI 10.1007/978-3-319-73356-2_1

Fjeld J, Achten N, Hilligoss H, Nagy A, Srikumar M (2020) Principled artificial intelligence: Mapping consensus in ethical and rights-based approaches to principles for ai. Berkman Klein Center Research Publication (2020-1), DOI 10.2139/ssrn.3518482

Fleischmann KR, Wallace WA (2005) A covenant with transparency: Opening the black box of models. Communications of the ACM 48(5):93–97, DOI 10.1145/1060710.1060715

Floridi L, Cowls J, Beltrametti M, Chatila R, Chazerand P, Dignum V, Luetge C, Madelin R, Pagallo U, Rossi F, Schafer B, Valcke P, Vayena E (2018) AI4People—an ethical framework for a good AI society: Opportunities, risks, principles, and recommendations. Minds and Machines (28):689–707, DOI 10.1007/s11023-018-9482-5

Foyer P (2015) Early experience, maternal care and behavioural test design: Effects on the temperament of military working dogs. PhD thesis, Linköping University Electronic Press, DOI 10.3384/diss.diva-122260

Franke U (2022) First-and second-level bias in automated decision-making. Philosophy & Technology 35(2):1–20, DOI 10.1007/s13347-022-00500-y

Frauenberger C (2019) Entanglement HCI the next wave? ACM Trans Comput-Hum Interact 27(1), DOI 10.1145/3364998

Friedman B, Nissenbaum H (1996) Bias in computer systems. ACM Transactions on Information Systems (TOIS) 14(3):330–347, DOI 10.1145/230538.230561

Gasser U, Almeida VA (2017) A layered model for AI governance. IEEE Internet Computing 21(6):58–62, DOI 10.1109/MIC.2017.4180835

GDPR (2016) Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing Directive 95/46/EC (General Data Protection Regulation). Official Journal of the European Union (OJ) L 119, 4.5. pp.1–88., URL `http://data.europa.eu/eli/reg/2016/679/oj`

Guidotti R, Monreale A, Ruggieri S, Turini F, Giannotti F, Pedreschi D (2018) A survey of methods for explaining black box models. ACM computing surveys (CSUR) 51(5):1–42, DOI 10.1145/3236009

Hamon R, Junklewitz H, Malgieri G, Hert PD, Beslay L, Sanchez I (2021) Impossible explanations? Beyond explainable AI in the GDPR from a Covid-19 use case scenario. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 549–559, DOI 10.1145/3442188.3445917

Hassenzahl M (2001) The effect of perceived hedonic quality on product appealingness. International Journal of Human-Computer Interaction 13(4):481–499, DOI 10.1207/S15327590IJHC1304_07

Hewett TT, Baecker R, Card S, Carey T, Gasen J, Mantei M, Perlman G, Strong G, Verplank W (1992) ACM SIGCHI curricula for human-computer interaction. ACM, DOI 10.1145/2594128

Hildebrandt M, Gutwirth S (2008) Profiling the European Citizen: Cross-Disciplinary Perspectives. Springer Science & Business Media, DOI 10.1007/978-1-4020-6914-7

Homewood S, Karlsson A, Vallgårda A (2020) Removal as a Method: A Fourth Wave HCI Approach to Understanding the Experience of Self-Tracking, Association for Computing Machinery, New York, NY, USA, p 1779–1791. DOI 10.1145/3357236.3395425

Hood C, Heald D (2006) Transparency: The Key to Better Governance? Proceedings of the British Academy, OUP/British Academy, URL `https://books.google.se/books?id=0B3rMQEACAAJ`

Hutchinson B, Smart A, Hanna A, Denton E, Greer C, Kjartansson O, Barnes P, Mitchell M (2021) Towards accountability for machine learning datasets: Practices from software engineering and infrastructure. In: Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, Association for Computing Machinery, New York, NY, USA, FAccT '21, p 560–575, DOI 10.1145/3442188.3445918

ISO (2018) ISO 9241-11:2018, Ergonomics of human-system interaction — Part 11: Usability: Definitions and concepts. ISO standard

Kang J, Hustvedt G (2014) Building trust between consumers and corporations: The role of consumer perceptions of transparency and social responsibility. Journal of Business Ethics 125(2):253–265, DOI 10.1007/s10551-013-1916-7

Keil FC (2006) Explanation and understanding. Annu Rev Psychol 57:227–254, DOI 10.1146/annurev.psych.57.102904.190100

Keyes O, Hoy J, Drouhard M (2019) Human-computer insurrection: Notes on an Anarchist HCI. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, CHI '19, p 1–13, DOI 10.1145/3290605.3300569

Kim S, Lee J (2012) E-participation, transparency, and trust in local government. Public Administration Review 72(6):819–828, DOI 10.1111/j.1540-6210.2012.02593.x

Kim SB, Kim DY (2017) Antecedents of corporate reputation in the hotel industry: The moderating role of transparency. Sustainability 9(6):951, DOI 10.3390/su9060951

Kizilcec RF (2016) How much information? Effects of transparency on trust in an algorithmic interface. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, pp 2390–2395, DOI 10.1145/2858036.2858402

Krebs LM, Alvarado Rodriguez OL, Dewitte P, Ausloos J, Geerts D, Naudts L, Verbert K (2019) Tell me what you know: GDPR implications on designing transparency and accountability for news recommender systems. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, New York, NY, USA, CHI EA '19, p 1–6, DOI 10.1145/3290607.3312808

Kwan D, Cysneiros LM, do Prado Leite JCS (2021) Towards achieving trust through transparency and ethics. In: 2021 IEEE 29th International Requirements Engineering Conference (RE), pp 82–93, DOI 10.1109/RE51729.2021.00015

Lakkaraju H, Bastani O (2020) "How do i fool you?" manipulating user trust via misleading black box explanations. In: Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, pp 79–85, DOI 10.1145/3375627.3375833

Langer M, Baum K, Hartmann K, Hessel S, Speith T, Wahl J (2021a) Explainability auditing for intelligent systems: A rationale for multi-disciplinary perspectives. In: 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), pp 164–168, DOI 10.1109/REW53955.2021.00030

Langer M, Oster D, Speith T, Hermanns H, Kästner L, Schmidt E, Sesing A, Baum K (2021b) What do we want from explainable artificial intelligence (XAI)? – a stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence 296:103473, DOI 10.1016/j.artint.2021.103473

Larsson L (2022) Allt går fel — därför störtdyker facebook. Dagens Nyheter (February 4, 22), URL `https://www.dn.se/ekonomi/linus-larsson-allt-gar-fel-darfor-stortdyker-facebook/`, accessed on March 3, 2022.

Larsson S (2020) On the governance of artificial intelligence through ethics guidelines. Asian Journal of Law and Society 7(3):437–451, DOI 10.1017/als.2020.19

Larsson S, Heintz F (2020) Transparency in artificial intelligence. Internet Policy Review 9(2), DOI 10.14763/2020.2.1469

Lessig L (1998) The New Chicago School. The Journal of Legal Studies 27(S2):661–691, DOI 10.1086/468039

Lessig L (1999) Code and Other Laws of Cyberspace. Basic Books, USA

Lessig L (2006) Code: Version 2.0. Basic Books, USA, URL `https://books.google.se/books?id=lmXIMZiU8yQC`

Lessig L (2009) Against transparency. URL `https://newrepublic.com/article/70097/against-transparency`, accessed on February 27, 2022.

Lindblad-Gidlund K, Ekelin A, Eriksén S, Ranerup A (2010) Förvaltning och medborgarskap i förändring. Studentlitteratur

London AJ (2019) Artificial intelligence and black-box medical decisions: Accuracy versus explainability. Hastings Center Report 49(1):15–21, DOI 10.1002/hast.973

Machuletz D, Böhme R (2020) Multiple purposes, multiple problems: A user study of consent dialogs after GDPR. In: Proceedings on Privacy Enhancing Technologies, vol 2, pp 481–498, DOI 10.48550/arXiv.1908.10048

Massey AK, Smith B, Otto PN, Antón AI (2011) Assessing the accuracy of legal implementation readiness decisions. In: 2011 IEEE 19th International Requirements Engineering Conference, pp 207–216, DOI 10.1109/RE.2011.6051661

McGregor L, Murray D, Ng V (2019) International human rights law as a framework for algorithmic accountability. International and Comparative Law Quarterly 68(2):309–343, DOI 10.1017/S0020589319000046

Meske C, Bunde E, Schneider J, Gersch M (2020) Explainable artificial intelligence: Objectives, stakeholders, and future research opportunities. Information Systems Management 0(0):1–11, DOI 10.1080/10580530.2020.1849465

Momen N, Hatamian M, Fritsch L (2019) Did app privacy improve after the GDPR? IEEE Security & Privacy 17(6):10–20, DOI 10.1109/MSEC.2019.2938445

Monett D, Lewis CWP, Thórisson KR, Bach J, Baldassarre G, Granato G, Berkeley ISN, Chollet F, Crosby M, Shevlin H, Fox J, Laird JE, Legg S, Lindes P, Mikolov T, Rapaport WJ, Rojas R, Rosa M, Stone P, Sutton RS, Yampolskiy RV, Wang P, Schank R, Sloman A, Winfield A (2020) Special issue "on defining artificial intelligence"—commentaries and author's response. Journal of Artificial General Intelligence 11(2):1–100, DOI 10.2478/jagi-2020-0003

Morey T, Forbath T, Schoop A (2015) Customer data: Designing for transparency and trust. Harvard Business Review 93(5):96–105

Noble SU (2018) Algorithms of Oppression: How Search Engines Reinforce Racism. NYU Press, DOI 10.18574/9781479833641

Norman D (1988) The Design of Everyday Things. Basic Books

Nouwens M, Liccardi I, Veale M, Karger D, Kagal L (2020) Dark patterns after the GDPR: Scraping consent pop-ups and demonstrating their influence. In: Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, pp 1–13, DOI 10.1145/3313831.3376321

Nozick R (1993) The nature of rationality. Princeton University Press

OECD (2017) Algorithms and collusion: Competition policy in the digital age. Tech. rep., OECD, URL `www.oecd.org/competition/algorithms-collusion-competition-policy-in-the-digital-age.html`

O'Neil C (2016) No safe zone – getting insurance. In: Weapons of math destruction: How big data increases inequality and threatens democracy, Broadway Books, pp 161–178

Park H, Blenkinsopp J (2011) The roles of transparency and trust in the relationship between corruption and citizen satisfaction. International Review of Administrative Sciences 77(2):254–274, DOI 10.1177/0020852311399230

Presthus W, Munkvold BE (2016) How to frame your contribution to knowledge? a guide for junior researchers in information systems. NOKOBIT - Norsk konferanse for organisasjoners bruk av informasjonsteknologi

Rader E, Cotter K, Cho J (2018) Explanations as Mechanisms for Supporting Algorithmic Transparency, Association for Computing Machinery, New York, NY, USA, p 1–13. DOI 10.1145/3173574.3173677

Rai A (2020) Explainable AI: from black box to glass box. Journal of the Academy of Marketing Science 48(0):137–141, DOI 10.1007/s11747-019-00710-5

Rebala G, Ravi A, Churiwala S (2019) Machine Learning Definition and Basics, Springer International Publishing, Cham, pp 1–17. DOI 10.1007/978-3-030-15729-6_1

Ribeiro M, Singh S, Guestrin C (2016) "Why should i trust you?": Explaining the predictions of any classifier. pp 97–101, DOI 10.18653/v1/N16-3020

Rogers SJ, Fang JH, Karr CL, Stanley DA (1992) Determination of lithology from well logs using a neural network. AAPG Bulletin 76(5):731–739, DOI 10.1306/BDFF88BC-1718-11D7-8645000102C1865D

Rossi F (2018) Building trust in artificial intelligence. Journal of International Affairs 72(1):127–134, DOI 10.2307/26588348

Rydenfält C, Persson J (2020) The usability and digitalization of healthcare: third-wave HCI meets first-wave challenges. XRDS: Crossroads, The ACM Magazine for Students 26(3):42–45, DOI 10.1145/3383386

Sadeghi M, Klös V, Vogelsang A (2021) Cases for explainable software systems: Characteristics and examples. In: 2021 IEEE 29th International Requirements Engineering Conference Workshops (REW), pp 181–187, DOI 10.1109/REW53955.2021.00033

Samek W, Binder A, Montavon G, Lapuschkin S, Müller KR (2016) Evaluating the visualization of what a deep neural network has learned. IEEE transactions on neural networks and learning systems 28(11):2660–2673, DOI 10.1109/TNNLS.2016.2599820

Sanchez-Rola I, Dell'Amico M, Kotzias P, Balzarotti D, Bilge L, Vervier PA, Santos I (2019) Can I opt out yet?: GDPR and the global illusion of cookie control. In: Proceedings of the 2019 ACM Asia conference on computer and communications security, pp 340–351, DOI 10.1145/3321705.3329806

Schneier B (2019) There's no good reason to trust blockchain technology. Wired Magazine Accessed on May 23, 2022.

Scott J (2004) Ethics, governance, trust, transparency and customer relations. The Geneva Papers on Risk and Insurance Issues and Practice 29(1):45–51, URL `https://www.jstor.org/stable/41952740`

Selbst AD, Powles J (2017) Meaningful information and the right to explanation. International Data Privacy Law 7(4):233–242, DOI 10.1093/idpl/ipx022

Siau K, Wang W (2018) Building trust in artificial intelligence, machine learning, and robotics. Cutter Business Technology Journal 31:47–53

Simmons A, Chappell S (1988) Artificial intelligence - definition and practice. IEEE Journal of Oceanic Engineering 13(2):14–42, DOI 10.1109/48.551

Simon HA (1971) Designing organizations for an information rich world. In: Greenberger M (ed) Computers, communications, and the public interest, Baltimore, pp 37–72

Sørum H, Presthus W (2020) Dude, where's my data? The GDPR in practice, from a consumer's point of view. Information Technology & People 34(3):912–929, DOI 10.1108/ITP-08-2019-0433

Speicher M (2015) What is usability? A characterization based on ISO 9241-11 and ISO/IEC 25010. arXiv preprint arXiv:150206792 DOI 10.48550/arXiv.1502.06792

Syrmoudis E, Mager S, Kuebler-Wachendorff S, Pizzinini P, Grossklags J, Kranz J (2021) Data portability between online services: An empirical analysis on the effectiveness of GDPR art. 20. In: Proceedings on Privacy Enhancing Technologies, vol 3, pp 351–372

Temme M (2017) Algorithms and transparency in view of the new General Data Protection Regulation. Eur Data Prot L Rev 3:473, DOI 10.21552/edpl/2017/4/9

TF (1766) Kongl. Maj:ts Nådige Förordning, Angående Skrif och Erna friheten. Leglislation passed in Stockholm, 2nd of December 1766

The Economist (2017) The world's most valuable resource is no longer oil, but data. The Economist (May 6), URL `https://www.economist.com/leaders/2017/05/06/the-worlds-most-valuable-resource-is-no-longer-oil-but-data`, accessed on May 23, 2022.

Turilli M, Floridi L (2009) The ethics of information transparency. Ethics and Information Technology 11(2):105–112, DOI 10.1007/s10676-009-9187-9

Tversky A, Kahneman D (1974) Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. science 185(4157):1124–1131

Velkova J, Kaun A (2021) Algorithmic resistance: media practices and the politics of repair. Information, Communication & Society 24(4):523–540, DOI 10.1080/1369118X.2019.1657162

Wachter S, Mittelstadt B, Floridi L (2017) Why a right to explanation of automated decision-making does not exist in the General Data Protection Regulation. International Data Privacy Law 7(2):76–99, DOI 10.1093/idpl/ipx005

Wang P (2019) On defining artificial intelligence. Journal of Artificial General Intelligence 10(2):1–37, DOI 10.2478/jagi-2019-0002

Wilson RA, Keil F (1998) The shadows and shallows of explanation. Minds and Machines 8(1):137–159, DOI 10.1023/A:1008259020140

Winner L (1980) Do artifacts have politics? Daedalus (Cambridge, Mass) 109(1):121–136

Wischmeyer T (2020) Artificial intelligence and transparency: Opening the black box. In: Regulating Artificial Intelligence, Springer, pp 75–101, DOI 10.1007/978-3-030-32361-5_4

Woodward J, Ross L (2021) Scientific Explanation. In: Zalta EN (ed) The Stanford Encyclopedia of Philosophy, Summer 2021 edn, Metaphysics Research Lab, Stanford University, URL `https://plato.stanford.edu/archives/sum2021/entries/scientific-explanation/`

Yanofsky NS (2011) Towards a definition of an algorithm. Journal of Logic and Computation 21(2):253–286, DOI 10.1093/logcom/exq016

Zerilli J, Knott A, Maclaurin J, Gavaghan C (2019) Transparency in algorithmic and human decision-making: Is there a double standard? Philosophy & Technology 32(4):661–683, DOI 10.1007/s13347-018-0330-6

Zuboff S (2018) The Age of Surveillance Capitalism: The Fight for a Human Future at the New Frontier of Power, 1st edn